Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY - ORGANIZED IN 1935

VOLUME 21
NUMBER 2
JUNE
1956

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

JUNE, 1956, VOLUME 21, NUMBER 2

Published by the Psychometric Society at 1407 Sherwood Avenue, Richmond 5, Virginia. Second-class mail privileges authorized at Richmond, Virginia. Editorial Office, Department of Psychology, The University of North Carolina, Chapel Hill, North Carolina.

Subscription Price: The regular subscription rate is $14.00 per volume. The subscriber receives each issue as it comes out, and, upon request, a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues but two are available. Back issues are $14.00 per volume (one set only) or $3.50 per issue, with a 20 per cent discount to Psychometric Society Members. Members of the Psychometric Society pay annual dues of $7.00, of which $6.30 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of $4.00, of which $3.60 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

LEONARD KOGAN
105 E. 22nd St., New York 10, N. Y.

Payments: All bills and orders are payable in advance.

Checks covering membership dues should be made payable to the *Psychometric Society*.

Checks covering regular subscriptions to *Psychometrika* (for non-members of the Psychometric Society) and back issue orders should be made payable to the *Psychometric Corporation*. All checks, notices of change of address, and business communications should be sent to

DAVID R. SAUNDERS, Treasurer, Psychometric Society and Psychometric Corporation
Educational Testing Service
P.O. Box 592
Princeton, New Jersey

Articles on the following subjects are published in *Psychometrika:*

(1) the development of quantitative rationale for the solution of psychological problems;

(2) general theoretical articles on quantitative methodology in the social and biological sciences;

(3) new mathematical and statistical techniques for the evaluation of psychological data;

(4) aids in the application of statistical techniques, such as nomographs, tables, work-sheet layouts, forms, and apparatus;

(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), in so far as articles of this type are available.

(*Continued on the back inside cover page*)

Psychometrika

CONTENTS

VOLUME TWENTY-ONE JUNE 1956 NUMBER 2

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

OPTIMAL TEST LENGTH FOR MAXIMUM ABSOLUTE PREDICTION*

PAUL HORST AND CHARLOTTE MACEWAN
UNIVERSITY OF WASHINGTON

The concepts of multiple differential prediction and multiple absolute prediction are developed in earlier papers (2, 3). The problem of determining the optimal distribution of testing time for multiple differential prediction has been previously considered (4). This paper develops an analogous procedure for multiple absolute prediction. A numerical example illustrating the procedure is presented. The mathematical rationale underlying the procedure is given.

I. *The Problem*

A technique was presented in (3) for selecting from a large number of predictor variables the subset of specified size which would have the highest absolute prediction efficiency for a given set of criterion variables. In (2), an analogous procedure was developed for selecting the subset which would most efficiently predict the multiple criteria differentially. In each of these cases, efficiency was defined in terms of the accuracy of prediction. In (2), differential prediction efficiency was defined in terms of the accuracy with which differences between all possible pairs of criterion measures could be predicted. An appropriate index of the prediction efficiency of a selected battery was shown to be the difference between the average variance of the predicted criteria and their average covariance. This index was designated by ϕ; the larger the value of ϕ, the greater the differential prediction efficiency of the battery.

In the case of multiple absolute prediction (3), efficiency was defined in terms of the accuracy with which all the criteria could be predicted, regardless of the extent to which the selected battery would differentiate among them. The index of absolute prediction efficiency of the selected battery was taken as the sum of the variances of the predicted criteria, regardless of their covariance; the larger this sum, designated by λ, the greater the absolute prediction efficiency of the battery.

*This research was carried out under Contract Nonr-477(08) between the University of Washington and the Office of Naval Research. The computations were carried out by Robert Dear and Donald Mills. Much credit is due the typist, Elizabeth Cross. Supervision of both computational and editorial activities was provided by William Clemans. To each of these able contributors we are deeply grateful.

For both types of multiple prediction it was assumed that intercorrelations between the potential predictors were available, as well as satisfactory estimates of the correlations between each criterion variable and the potential predictors. It was further assumed that predictors and criteria were in standard measures, and that the predicted criteria were the least square estimates. As was pointed out in (3), the essential difference between multiple absolute and multiple differential prediction is that the respective sets of predictors selected will differ. The two sets may show varying degrees of overlap; the extent of overlap depends upon the degree of correlation among the criterion variables and upon the original group of potential predictors from which the two subsets were selected.

Both methods referred to tacitly assume that all potential predictors take the same amount of administration time, so that all subsets of the same size would take equal administration time. This will not usually be the case. The problem may be approached in a more general way by starting with a given battery of predictor variables and determining how, for a specified over-all testing time, the individual test time-limits should be altered in order to maximize the index of absolute prediction efficiency.

For the case of a single criterion, a method is available (1) for determining optimal distribution of over-all administration time for a given battery of predictors. In (4) the method is modified and generalized for differential prediction involving multiple criteria. In this article the procedure developed in (1) is extended to the case of multiple absolute prediction.

As published, the method presented in (1) for solving for optimal test length assumes that the regression weights for the tests of optimal length are all positive. Otherwise the optimal solution could lead to the unacceptable solution where some of the optimal test lengths are negative. The extension of the technique in this article provides a computational solution which cannot yield negative test length. Since the method in (1) is a special case of the technique presented in the present article the more general method may also be used to avoid negative test lengths for the case of a single criterion. However, an iterative solution is required, whereas in the former case an exact solution was indicated.

In the present paper, as in the cases previously presented, it is assumed that intercorrelation, validity, and reliability data are available for predictors of arbitrary length. Testing time is taken to be the time actually allotted the examinee for taking the test. Any alteration in testing time implies a corresponding alteration in the number of items. Consequently the terms "testing time" and "test length" are used synonymously.

The method will first be described and illustrated by a numerical example. Following this, the mathematical rationale will be presented.

II. *Numerical Example*

The data used for this example are the same as those presented in (4). The predictor variables are:

(1) Guilford-Zimmerman Aptitude Survey Part I, Verbal Comprehension

(2) Guilford-Zimmerman Aptitude Survey Part III, Numerical Operations

(3) Guilford-Zimmerman Aptitude Survey Part VII, Mechanical Knowledge

(4) A. C. E. Psychological Examination, Quantitative Reasoning

(5) A. C. E. Psychological Examination, Linguistic Reasoning

(6) Cooperative English Test (Form OM), Usage

The matrix of test intercorrelations with reliabilities in the diagonal is given in Table 1. The criterion variables are grade point averages in each of ten college areas. The matrix of validity coefficients is given in Table 2.

It is evident that the correlations of variable 3 with the criterion are all small, four of them being negative. In general the chief justification for including such a variable would be that it might serve as a suppressor variable, i.e., a variable which suppresses invalid systematic variance in the predictor variables.

The original test lengths may be seen in row 1 of Table 3. The over-all testing time for the tests of arbitrary length is 142 minutes. We assume that this time is to be cut in half so that the over-all testing time is 71 minutes. The problem is to determine the time to be allotted to each test so as to maximize the index of absolute prediction efficiency.

The traditional assumptions are used here as in (1) with respect to the effect of test length on correlation, and will not be repeated. As in (4), the method for solving for the new test lengths involves a series of successive approximations. The computational procedure to be described consists of the same sequence of operations as those given in (4), the difference being that in the current procedure the matrix of validity coefficients is used, whereas in (4) the deviation form of these coefficients was required.

1. The first computational step is the calculation of the elements for a diagonal matrix Δ, seen in Table 3, row 3, labelled $1'\Delta$. Row 1 of this table gives the original lengths of tests; the elements in row 2 were obtained by subtracting the reliability of the indicated test from unity. The elements of Δ in row 3 are obtained by multiplying the element in row 1 by the corresponding element in row 2. Thus for the first element we have $\Delta_1 = 25(.080) = 2.000$.

2. A first approximation is now required for the altered test lengths. We assume the new test lengths to be proportional to the original test lengths.

TABLE 1

R Matrix of Predictor Intercorrelations with Reliabilities Substituted for Unities in the Diagonal:

$R = r - D_u$

		1	2	3	4	5	6	Σ
1	G-Z I	.920	.159	.152	.281	.763	.515	2.790
2	G-Z III	.159	.920	.003	.369	.292	.243	1.986
3	G-Z VII	.152	.003	.920	.200	.142	-.150	1.267
4	A.C.E.-Q	.281	.369	.200	.820	.549	.426	2.645
5	A.C.E.-L	.763	.292	.142	.549	.830	.628	3.204
6	English	.515	.243	-.150	.426	.628	.860	2.522
Σ		2.790	1.986	1.267	2.645	3.204	2.522	14.414

TABLE 2

The r'_c Matrix of Validity Coefficients

		1 G-Z I	2 G-Z III	3 G-ZVII	4 ACE-Q	5 ACE-L	6 English	Σ
1	Anthropology	.370	.177	.091	.294	.341	.357	1.630
2	Chemistry	.317	.274	.016	.309	.364	.399	1.679
3	Economics	.339	.211	.008	.241	.334	.323	1.456
4	English	.526	.247	-.075	.262	.488	.524	1.972
5	Foreign Lang.	.295	.287	-.156	.200	.232	.426	1.284
6	Geology	.184	.140	.094	.170	.229	.214	1.031
7	History	.379	.169	-.001	.182	.373	.336	1.438
8	Mathematics	.287	.348	-.088	.350	.336	.401	1.634
9	Psychology	.440	.170	.096	.285	.409	.403	1.803
10	Zoology	.336	.216	.031	.318	.345	.351	1.597
Σ		3.473	2.239	.016	2.611	3.451	3.734	15.524

TABLE 3

Computation of $1'\Delta$, $1'D_{b_1}^{-1}$ and $1'\Delta D_{b_1}^{-1}$

First approximation: $1'D_{b_1} = \frac{1}{2}1'D_a$

		1	2	3	4	5	6	Ck	Σ
1	$1'D_a$	25.0	9.0	30.0	23.0	15.0	40.0		142.0
2	$1'D_u$	.080	.080	.080	.180	.170	.140		
3	$1'\Delta$	2.000	.720	2.400	4.140	2.550	5.600		17.410
4	$1'D_{b_1}=.5(1'D_a)$	12.5	4.5	15.0	11.5	7.5	20.0	71.0	71.0
5	$1'D_{b_1}-1$	.080	.222	.067	.087	.133	.050		
6	$1'\Delta D_{b_1}^{-1}$	.160	.160	.161	.360	.339	.280		1.460

Therefore, as a first approximation to the new test lengths we take one half the original test lengths. Row 4 in Table 3 is obtained as one half of row 1.

3. Calculate the reciprocal for each of the D_b elements. These are shown in row 5 of Table 3.

4. Multiply each Δ value in row 3 of Table 3 by the corresponding value in row 5. The products are entered in row 6 of Table 3. For example, the first element is .160 = 2.000(.080).

5. Next, the elements calculated in step 4 are added to the corresponding diagonal elements of Table 1, and the table is copied into the upper left quadrant of Table 4. The first diagonal element is 1.080 = .160 + .920. Note that the elements below the diagonal are not copied in. The upper right section of Table 4 is the matrix r_e , the transpose of Table 2.

6. We next calculate a matrix L_1 by premultiplying the matrix r_e by the inverse of the matrix in the upper left quadrant of Table 4. The computations of the forward solution are given in the two lower quadrants of Table 4 and in Table 5. The back solution is given in Table 6, in which the transpose of the matrix L_1 appears in the upper left corner. The procedure for multiplying a matrix by the inverse of a symmetric matrix is outlined in (**5**).

7. The second approximation to the new test lengths is computed in the lower section of Table 6 as follows:

Row *a* consists of the sums of squares of the column elements of the L_1' matrix. For example, the first element of row *a*, namely, .441, is the sum of squares of the first ten elements in column 1 of Table 6. Row *b* is copied from row 3 of Table 3. Row *c* consists of products of corresponding elements in rows *a* and *b*. For example, for the first element, .882 = .441(2.000).

Row *d* is obtained by taking the square root of the corresponding element in row *c*. For example, .939 = $\sqrt{.882}$. Row *e*, a check upon the computation of row *d*, consists of the squares of corresponding elements in row *d*. Row *e*, then, should be the same as row *c*. The value computed to the right of row *e*, and labelled s_1 , is obtained by dividing the new over-all testing time, 71 minutes, by 3.930, the sum of elements in row *d*. Thus the value of s_1 is 18.066.

Row *f* is obtained by multiplying each element of row *d*, including the summation element, by s_1 . For example, the first element obtained is 16.964 = .939(18.066). This row gives the second approximation to the new test lengths and its sum should equal the new over-all testing time.

Row *g* is obtained by dividing each element in row *b* by the corresponding element in row *f*. For example, the first element is .118 = 2.000/16.964.

Row *h* is obtained by adding the elements in row *g* to the corresponding reliabilities as given in the diagonal of Table 1. For example, the first element in row *h* is .118 + .920 = 1.038.

8. A new L_2 matrix is computed by repeating steps 5 and 6, and using the elements of row *h* of Table 6 in the diagonal positions of Table 1. The L_2 matrix may be seen in transposed form in Table 7, rows 1 through 10.

TABLE 4

Computation of $L_1 = (R+\Delta D_{b_1}^{-1})^{-1} r_c$ Forward Solution

Recip	i	1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	7B	8B	9B	10B	C	S
	1A	1.080	.159	.152	.281	.763	.515	.370	.317	.339	.526	.295	.184	.379	.287	.440	.336	6.423	
	2A		1.080	.003	.369	.292	.243	.177	.274	.211	.247	.287	.140	.169	.348	.170	.216	4.385	
	3A			1.081	.200	.142	-.150	.091	.016	.008	-.075	-.156	.094	-.001	-.088	.096	.031	1.444	
	4A				1.180	.549	.426	.294	.309	.241	.262	.200	.170	.182	.350	.285	.318	5.616	
	5A					1.169	.628	.341	.364	.334	.488	.232	.229	.373	.336	.409	.345	6.994	
	6A						1.140	.357	.399	.323	.524	.426	.214	.336	.401	.403	.351	6.536	
	C	2.950	2.146	1.428	3.005	3.543	2.802	1.630	1.679	1.456	1.972	1.284	1031	1.438	1.634	1.803	1.597	31.398	31.398
.926	1B	1.080	.159	.152	.281	.763	.515	.370	.317	.339	.526	.295	.184	.379	.287	.440	.336	6.423	6.423
.946	2B		1.057	-.019	.328	.180	.167	.123	.227	.161	.170	.244	.113	.113	.306	.105	.167	3.441	3.442
.944	3B			1.059	.166	.038	-.220	.041	-.025	-.037	-.146	-.193	.070	-.052	-.123	.036	-.013	.600	.601
1.021	4B				.979	.289	.275	.153	.160	.109	.095	.078	.076	.057	.200	.132	.181	2.785	2.784
1.953	5B					.512	.162	.012	.055	.036	.064	-.034	.055	.071	.026	.040	.026	1.025	1.025
1.441	6B						.694	.123	.144	.086	.169	.195	.084	.088	.126	.134	.103	1.947	1.946

TABLE 5

Computation of $L_1 = (R+\Delta D_{b_1}^{-1})^{-1} r_c$ (continued)

i	1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	7B	8B	9B	10B	C	S
1A	1.000																	
2A		1.000																
3A			1.000															
4A				1.000														
5A					1.000													
6A						1.000												
1B	-1.000	-.147	-.141	-.260	-.707	-.477	-.343	-.294	-.314	-.487	-.273	-.170	-.351	-.266	-.407	-.311	-5.948	-5.948
2B		-1.000	.018	-.310	-.170	-.158	-.116	-.215	-.152	-.161	-.231	-.107	-.107	-.289	-.099	-.158	-3.255	-3.255
3B			-1.000	-.157	-.036	.208	-.039	.024	.035	.138	.182	-.066	.049	.116	-.034	.012	- .566	- .568
4B				-1.000	-.295	-.281	-.156	-.163	-.111	-.097	-.080	-.078	-.058	-.204	-.135	-.185	-2.843	-2.843
5B					-1.000	-.316	-.023	-.107	-.070	-.125	.066	-.107	-.139	-.051	-.078	-.051	-2.002	-2.001
6B						-1.000	-.177	-.208	-.124	-.244	-.281	-.121	-.127	-.182	-.193	-.148	-2.806	-2.805

TABLE 6

Computation of $L_1 = (R+\Delta D_{b_1}^{-1})^{-1} r_c$ Back Solution

1		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	7B	8B	9B	10B	Check	
1		.235	.059	.059	.116	-.033	.177	-1										.002	0
2		.120	.147	.003	.092	.041	.208		-1									-.001	0
3		.203	.106	-.021	.067	.031	.124			-1								-.001	0
4		.330	.108	-.091	.014	.048	.244				-1							-.001	0
5		.225	.196	-.125	.047	-.155	.281					-1						.001	0
6		.035	.070	.085	.024	.069	.121						-1					-.001	0
7		.215	.072	-.025	-.007	.099	.127							-1				.001	0
8		.127	.212	-.102	.155	-.007	.182								-1			.001	0
9		.268	.043	.060	.076	.017	.193									-1		-.005	0
10		.188	.090	-.004	.142	.004	.148										-1	.001	0
a	$1'D_{L_1L_1'}$	.441	.151	.050	.079	.045	.352	Σ	Check										
b	$1'\Delta$	2.000	.720	2.400	4.140	2.550	5.600												
c	$1'D_{L_1L_1'}\Delta$	.882	.109	.120	.327	.115	1.971												
d	$1'(D_{L_1L_1'}\Delta)^{\frac{1}{2}}$	.939	.330	.346	.572	.339	1.404	3.930	3.930										
e	$1'(D_{L_1L_1'}\Delta)^{\frac{1}{2}}(D_{L_1L_1'}\Delta)^{\frac{1}{2}}$	.882	.109	.120	.327	.115	1.971												
f	$1'D_{b_2} = 1'(D_{L_1L_1'}\Delta)^{\frac{1}{2}}s_1$	16.964	5.962	6.251	10.334	6.124	25.365	71.000	71.000										
g	$1'\Delta D_{b_2}^{-1}$	.118	.121	.384	.400	.416	.221	1.660											
h	$1'D_R + 1'\Delta D_{b_2}^{-1}$	1.038	1.041	1.304	1.220	1.246	1.081	6.930	6.930										

$$s_1 = \frac{T}{1'(D_{L_1L_1'}\Delta)^{\frac{1}{2}}1} = \frac{71}{3.930} = 18.066$$

TABLE 7

Computation of $L_2 = (R + \Delta D_{b_2}^{-1})^{-1} r_c$ Back Solution

i		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	7B	8B	9B	10B	Check	
1		.247	.061	.049	.110	-.040	.186	-1										-.002	0
2		.126	.153	.008	.085	.026	.227		-1									-.001	0
3		.216	.111	-.015	.062	.016	.135			-1								-.002	0
4		.350	.113	-.071	.008	.020	.268				-1							-.001	0
5		.218	.203	-.099	.030	-.151	.307					-1						.001	0
6		.044	.073	.071	.026	.056	.127						-1					-.002	0
7		.235	.076	-.019	-.006	.072	.140							-1				.001	0
8		.125	.223	-.079	.139	-.014	.204								-1			.001	0
9		.288	.046	.052	.073	.002	.202									-1		.001	0
10		.199	.095	-.001	.134	-.007	.160										-1	.002	0
a	$1'D_{L_2L_2'}$	.489	.165	.032	.067	.034	.414		Σ	Check									
b	$1'\Delta$	2.000	.720	2.400	4.140	2.550	5.600												
c	$1'D_{L_2L_2'}\Delta$	.978	.119	.077	.277	.087	2.318		3.856	3.856									
d	$1'(D_{L_2L_2'}\Delta)^{\frac{1}{2}}$	.989	.345	.277	.526	.295	1.522		3.954										
e	$1'(D_{L_2L_2'}\Delta)(D_{L_2L_2'}\Delta)^{\frac{1}{2}}$	.978	.119	.077	.277	.087	2.316												
f	$1'D_{b_3} = 1'(D_{L_2L_2'}\Delta)^{\frac{1}{2}}s_2$	17.758	6.195	4.974	9.445	5.297	27.329		70.999	70.998									
g	$1'\Delta D_{b_3}^{-1}$	.113	.116	.482	.438	.481	.205		1.835										
h	$1'D_R + 1'\Delta D_{b_3}^{-1}$	1.033	1.036	1.402	1.258	1.311	1.065		7.105		7.105								

$$s_2 = \frac{T}{1'(D_{L_2L_2'}\Delta)^{\frac{1}{2}}1} = \frac{71}{3.954} = 17.956$$

9. Step 8 is repeated in rows *a* through *h* in Table 7, where a third approximation to the new test lengths is seen in row *f*.

Steps 5, 6, and 7 were again repeated to obtain a fourth approximation to the new test lengths. The values so obtained are shown, together with those of the preceding approximations, in Table 8. The vector of altered test lengths has not yet completely stabilized, but may be considered sufficiently so for all practical purposes.

10. The index of absolute prediction efficiency, λ, is computed as follows:

(*a*) To obtain the index λ, corresponding to the first approximation to optimal test lengths, calculate the product of each value in the L_1 matrix by the corresponding value in the validity matrix seen in Table 2, and sum the products. Thus the value of λ_1, seen as the first entry in the column to the far right in Table 8, is 2.053.

TABLE 8

Successive Approximations to $1'D_b$, for $T_1 = \frac{1}{2}T_0 = \frac{142}{2}$

Approx'n		1	2	3	4	5	6	Σ	Value of λ for successive values in L	
(.5) 1 D_a :	1	12.50	4.50	15.00	11.50	7.50	20.00	71.00	L_1	2.053
	2	16.96	5.96	6.25	10.33	6.12	25.36	70.98	L_2	2.086
	3	17.76	6.20	4.97	9.44	5.30	27.33	71.00	L_3	2.089
	4	17.90	6.30	4.61	9.10	4.81	28.28	71.00	L_4	2.090

(*b*) For λ_2, use the L_2 matrix instead of L_1, and repeat the procedure described in (*a*).

(*c*) To obtain any subsequent index, λ_i, substitute the L_i matrix for the L_1 matrix, while following the procedure indicated in (*a*).

The values of λ show only a very small increase in this particular illustration. From the initial value of 2.053 to a value of 2.090 corresponding to the fourth approximation, the increase is .037 or less than two per cent, although some of the test lengths are altered to a considerable extent. In the case of differential prediction (**4**), the increase in the efficiency index, ϕ, based on the same original data, was relatively larger, four per cent, for the fourth approximation to optimal test lengths.

Computations were also carried out with the original over-all testing time unchanged, and with the over-all testing time doubled. Three iterations were calculated for each of these conditions. The successive approximations to optimal test length, with the corresponding λ values, may be seen in Tables 9 and 10, respectively. Under these conditions, also, the increase in λ is small, although the alteration of some of the test lengths is relatively greater than that found in the first illustration. For unchanged over-all testing time, the increase in λ from the first to the third approximation is 1.5 per cent; for

TABLE 9

Successive Approximations to $1'D_b$, for $T_1 = T_0 = 142$

Approx'n	1	2	3	4	5	6	Σ	Value of λ for Successive Values in L	
(1.0) $1'D_a$: 1	25.00	9.00	30.00	23.00	15.00	40.00	142.00	L_1	2.203
2	32.54	10.00	10.45	21.42	18.66	48.93	142.00	L_2	2.230
3	32.87	9.70	8.21	20.21	21.61	49.40	142.00	L_3	2.234

TABLE 10

Successive Approximations to $1'D_b$, for $T_1 = 2T_0 = 2(142)$

Approx'n	1	2	3	4	5	6	Σ	Value of λ for Successive Values in L	
(2.0) $1'D_a$: 1	50.00	18.00	60.00	46.00	30.00	80.00	284.00	L_1	2.332
2	63.21	15.81	16.23	43.98	56.11	88.66	284.00	L_2	2.373
3	64.22	13.98	12.21	43.99	67.65	81.96	284.01	L_3	2.375

double the length of the original over-all testing time, the corresponding increase in λ is roughly two per cent. In the case of differential prediction (4), the corresponding increases in ϕ for three iterations were seven per cent and ten per cent, respectively.

It is obvious that for this particular example the prediction efficiency does not seem to be much improved by using optimal test lengths. Considerably more research may be required to indicate under what conditions, if any, prediction efficiency may be expected to vary appreciably as a function of variation in relative test length.

III. *Mathematical Derivation*

In (1) a procedure was developed for redistributing the specified over-all testing time for a battery of tests in such a way as to obtain optimal prediction of a single criterion variable. In this report, the procedure is extended to the problem of optimal prediction of multiple criteria. Let

$M \equiv$ the number of cases.
$n \equiv$ the number of predictors.
$N \equiv$ the number of criteria.
$Z \equiv$ an $(M \times n)$ matrix of test scores in a battery of altered length with the elements of Z of the form $(z_{ij} - \bar{z}_j)/(\sigma_{z_j}\sqrt{M})$.
$W \equiv$ an $(M \times N)$ matrix of criterion scores whose elements are deviate scores of the form $(w_{ij} - \bar{w}_j)/(\sigma_{w_j}\sqrt{M})$.
$B \equiv$ an $(n \times N)$ matrix of regression coefficients for estimating W from Z.
$r \equiv$ an $(n \times n)$ matrix of intercorrelations of tests of original lengths.
$\rho \equiv$ an $(n \times n)$ matrix of intercorrelations of tests of altered lengths.

$r_c \equiv$ an $(n \times N)$ matrix of validity coefficients for the tests of original lengths.
$\rho_c \equiv$ an $(n \times N)$ matrix of validity coefficients for the tests of altered lengths.
$D_a \equiv$ an $(n \times n)$ diagonal matrix of original test lengths.
$D_b \equiv$ an $(n \times n)$ diagonal matrix of altered test lengths.
$D_e = D_b D_a^{-1} \equiv$ the ratio of altered to original test lengths.
$D_{r_{ii}} \equiv$ the $(n \times n)$ diagonal matrix of reliability coefficients for the tests of original lengths.

The index of absolute prediction efficiency as defined in (3) is given by

$$\lambda = \text{tr } C,$$

where C is the covariance matrix of predicted criterion scores in standard measure. Let

$$\delta = [I + (D_e - I)D_{r_{ii}}]D_e^{-1}. \tag{1}$$

Let

$$\epsilon = (ZB - W). \tag{2}$$

We wish to minimize the trace of $\epsilon'\epsilon$ with the constraining condition, $1'D_b1 = T$, where T represents the new over-all testing time and 1 is a column vector of all unit elements. From (2),

$$\epsilon'\epsilon = B'Z'ZB - B'Z'W - W'ZB + W'W. \tag{3}$$

From the definitions above,

$$Z'Z = \rho, \tag{4}$$

$$Z'W = \rho_c . \tag{5}$$

From (3), (4), (5)

$$\epsilon'\epsilon = B'\rho B - B'\rho_c - \rho_c'B + W'W. \tag{6}$$

Let

$$\psi = \text{tr } \epsilon'\epsilon + \lambda 1'D_b1, \tag{7}$$

where λ is a Lagrangian multiplier. From (6) and (7)

$$\psi = \text{tr } (B'\rho B - B'\rho_c - \rho_c'B) + N + \lambda 1'D_b1. \tag{8}$$

In (1) it is shown, for the case of a single criterion variable, that

$$\rho_c = \delta^{-1/2}r_c , \tag{9}$$

a relationship readily seen to hold for the case of multiple criteria. It was also shown in (1) that

$$\rho = \delta^{-1/2}(r - D_u + D_uD_aD_b^{-1})\delta^{-1/2}, \tag{10}$$

where D_u is defined as $I - D_{r_{ii}}$, a diagonal matrix of test unreliability coefficients. Substituting (9) and (10) in (8),

$$\psi = \text{tr}\,[B'\delta^{-1/2}(r - D_u + D_u D_u D_b^{-1})\delta^{-1/2}B - B'\delta^{1/2} - r_c - r_c'\delta^{-1/2}B] + N + \lambda 1' D_b 1. \tag{11}$$

Let

$$B = \delta^{1/2}L, \tag{12}$$

$$r - D_u = R, \tag{13}$$

$$D_u D_u = \Delta. \tag{14}$$

Then (11) becomes

$$\psi = \text{tr}\,[L'(R + \Delta D_b^{-1})L - L'r_c - r_c'L] + N + \lambda 1' D_b 1. \tag{15}$$

The unknowns on the right side of (15) are L, D_b, and λ. Differentiating (15) with respect to row vectors of L', and setting this derivative equal to zero,

$$\frac{\partial \psi}{\partial L'} = (R + \Delta D_b^{-1})L - r_c = 0,$$

or

$$(R + \Delta D_b^{-1})L = r_c\,. \tag{16}$$

Differentiating (15) with respect to D_b and setting this derivative equal to zero,

$$\frac{\partial \psi}{\partial D_b} = \lambda I - D_{LL'}\Delta D_b^{-2} = 0, \tag{17}$$

where $D_{LL'}$ is a diagonal matrix whose non-zero elements are the diagonal elements of LL'. From (17)

$$D_b = (D_{LL'}\Delta)^{1/2}/\lambda^{1/2}. \tag{18}$$

From (18)

$$1'D_b 1 = 1'(D_{LL'}\Delta)^{1/2}1/\lambda^{1/2}, \tag{19}$$

or, since we have the constraining condition, $1'D_b 1 = T$,

$$\lambda^{1/2} = 1'(D_{LL'}\Delta)^{1/2}1/T. \tag{20}$$

Substituting (20) in (18), we obtain

$$D_b = \frac{(D_{LL'}\Delta)^{1/2}T}{1'(D_{LL'}\Delta)^{1/2}1}. \tag{21}$$

From equations (16) and (21), the formulas are derived for solving for D_b by a series of successive approximations. From (16) we obtain

$$L = (R + \Delta D_b^{-1})^{-1} r_e \,. \tag{22}$$

Let

$$L_i = (R + \Delta D_{b_i}^{-1})^{-1} r_e \tag{23}$$

where

$$D_{b_1} = \frac{D_e T}{1' D_e 1} \,, \tag{24}$$

and

$$D_{b_{i+1}} = \frac{(D_{L_i L_i'} \Delta)^{1/2} T}{1' (D_{L_i L_i'} \Delta)^{1/2} 1} \,. \tag{25}$$

The first approximation to D_b is indicated by (24). The second and all subsequent approximations to D_b may be obtained by an iterative procedure based on (23) and (25). Thus, successive approximations to L_i and $D_{b_{i+1}}$ may be computed until D_b stabilizes satisfactorily.

The regression vectors for the tests of optimal length will be given by

$$B = \rho^{-1} \rho_e \,. \tag{26}$$

From (9), (10), (13), and (14)

$$B = \delta^{1/2} (R + \Delta D_b^{-1})^{-1} r_e \,, \tag{27}$$

and from (22) and (27),

$$B = \delta^{1/2} L, \tag{28}$$

where L has been stabilized through successive approximations.

Furthermore, it can be shown that the index of absolute prediction efficiency, λ, as defined in (3) is given by

$$\lambda = \operatorname{tr} L' r_e \,. \tag{29}$$

It may be that one or more elements of $D_{b_{i+1}}$, as given in (25), may approach zero as i increases. In this case it would be better to write (23) in the form

$$L_i = D_{b_i}^{1/2} (D_{b_i}^{1/2} R D_{b_i}^{1/2} + \Delta)^{-1} D_{b_i}^{1/2} r_e \,. \tag{30}$$

Although the computation of successive L_i matrices by means of (30) would be computationally more laborious than with (23), it would avoid difficulties resulting from one or more near vanishing elements of D_b.

The computational procedure presented in Section II is related to the mathematical derivation as follows:

Table 1 is based on equation (13).

Step 1 is based on equation (14).

Step 2 is given by equation (24).

Step 3 consists of calculating $D_{b_1}^{-1}$ from D_{b_1} .

Step 4 consists of calculating ΔD_b^{-1} .

Step 5 consists of calculating the matrix within the parentheses in equation (23) for the case $i = 1$.

Step 6 consists of the computation of the matrix L_1 from equation (23).

Step 7 consists of calculating D_{b_2} from equation (25) for the case $i = 1$.

Step 8 consists of the computation of the matrix L_2 from equation (23).

Step 9 consists of calculating D_{b_2} from equation (25).

In general, successive approximations to L and D_b are obtained by repeating steps 6 and 7 for successive values of i in equations (23) and (25).

Step 10 follows the procedure indicated by equation (29) to obtain successive values of λ.

REFERENCES

1. Horst, P. Determination of optimal test length to maximize the multiple correlation. *Psychometrika*, 1949, **14**, 79–88.
2. Horst, P. A technique for the development of a differential prediction battery. *Psychol. Monogr.*, 1954, **68**, No. 9 (Whole No. 380).
3. Horst, P. A technique for the development of a multiple absolute prediction battery. *Psychol. Monogr.*, 1955, **69**, No. 5 (Whole No. 390).
4. Horst, P. Optimal test length for maximum differential prediction. *Psychometrika*, 1956, **21**, 51–66.
5. Horst, P. Servant of the human sciences, Chapt. 21, Section 7. Unpublished manuscript. Division of Counseling and Testing Services, University of Washington.

Manuscript received 3/2/55

Revised manuscript received 4/28/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

A LEAST SQUARES SOLUTION FOR PAIRED COMPARISONS WITH INCOMPLETE DATA*†

HAROLD GULLIKSEN

PRINCETON UNIVERSITY

AND

EDUCATIONAL TESTING SERVICE

A precise and rapid procedure has been devised for dealing with a matrix of incomplete data in paired comparisons. This method should increase the general applicability of paired comparisons since experiments involving large numbers of stimuli may now be shortened to feasible experimental proportions. Also, we may now use sets of stimuli which cover a wide range, resulting in a considerable number of 100 per cent vs. 0 per cent judgments, and still give a precise solution depending equally on each of the observations.

In scaling by paired comparisons, many cases arise where one does not have complete data. In any problem where the range of the set of stimuli is great in relation to the discriminal dispersion, there will be no usable data for the extreme comparisons. This is usually the case in dealing with construction of scales for various sensory areas, such as brightness, hue, pitch, or loudness. In other cases the experiment is made less laborious for the subject by not requiring all comparisons. If one is interested in studies of value judgments (6) where composite objects are used, such as (*a* and *b* vs. *c*) or (*a* and *b* vs. *c* and *d*), then it is also highly desirable not to include all possible comparisons. For example, it may be well to omit comparison (*a* and *b* vs. *b*) or the comparison (*a* and *b* vs. *a* and *c*). Such situations might arise in studying preferences for various foods, gifts, individuals, activities, or goals. Any of the types of situations indicated above may give rise to a matrix of incomplete data for which a reasonably precise solution is desired. A solution for paired comparisons with incomplete data for the case in which the correlations are equal and the ratios of the discriminal dispersions are known is presented here. The general usability of paired comparisons, especially in fields such as sensation and value judgments, will be

*This study was supported in part by the Office of Naval Research Contract N6onr-270-20 with Princeton University and by the National Science Foundation Grant G-642. The opinions expressed are, of course, those of the author and do not represent attitudes or policies of the Office of Naval Research or of the National Science Foundation.

†The author wishes to acknowledge helpful suggestions and comments received from Frederic M. Lord, Warren S. Torgerson, and Ledyard R Tucker. Thanks are also due to Mrs. Gertrude Diederich for some of the tabulating and computing for the illustrative problem.

greatly enhanced by a precise method for utilizing all the available data when dealing with an incomplete matrix.

The procedure to be presented follows the general method for handling incomplete data outlined by Horst (**2**, pp. 419-430), Kempthorne (**3**), or Kendall (**4**). The least squares solution for paired comparisons, when correlations are equal and ratios of discriminal dispersions are known, has been indicated by Mosteller (**5**). The problem may be stated in the following manner for the case in which complete data are available:

For the case of errorless data let S_i^* , S_j^* $(i = 1, \cdots, n; j = 1, \cdots, n)$ be the scale values of the stimuli. Define the distance $D_{ij}^* \equiv S_i^* - S_j^*$. From the values of p_{ij} (the experimental proportions of judgments $i > j$) we use a normal curve or some other assumption to determine values of D_{ij}, which are taken as estimates of D_{ij}^* .

The scaling problem is to determine values S_i , S_j $(i = 1, \cdots, n; j = 1, \cdots, n)$ such that $\sum_{i=1}^{n} \sum_{j=1}^{n} (D_{ij} - S_i + S_j)^2$ is a minimum. The subscripts i and j are alternative subscripts each designating the stimuli from 1 to n.

Least Squares Solution for Incomplete Data

For the case in which only incomplete data are available we may write

$$Q = \sum_{i,j}^{n_{ij}} (D_{ij} - S_i + S_j)^2, \tag{1}$$

using n_{ij} to indicate that the summation is over the available data. The scaling problem is to determine S_i and S_j $(i, j = 1, \cdots, n)$ such that Q is a minimum. It should be noted that since the matrix of D_{ij} for complete data is skew symmetric the matrix of D_{ij} for incomplete data will also be skew symmetric.

To determine the unknown S values that will make Q a minimum, take the partial derivative of Q with respect to S_i , giving

$$\frac{1}{2} \frac{\partial Q}{\partial S_i} = -\sum_{j}^{n_{ij}} (D_{ij} - S_i + S_j) + \sum_{j}^{n_{ij}} (-D_{ij} + S_i - S_j). \tag{2}$$

One term represents the partial derivative for the *row* in which S_i occurs in each cell, and the other term represents the partial derivative for the *column* of entries each of which contains S_i . The first term is identical with the second so we may write

$$\frac{1}{2} \frac{\partial Q}{\partial S_i} = 2 \sum_{j}^{n_{ij}} (S_i - S_j - D_{ij}). \tag{3}$$

If n_i represents the number of observations in row (or column) i, then simplifying and setting the partial derivative equal to zero gives

$$n_i S_i - \sum_{j}^{n_i} S_j - \sum_{j}^{n_i} D_{ij} = 0 \qquad (i = 1, \cdots, n). \tag{4}$$

This set of n equations may be expressed in matrix notation. Let

$S \equiv$ the column vector of elements S_i .

$1 \equiv$ a column vector of 1's.

$D \equiv$ the matrix with elements D_{ij} wherever observations occur, zero for each missing observation, and zeroes in the principal diagonal. An illustrative D matrix is shown in Table 3 provided a zero is put in for each missing entry.

$M \equiv$ a matrix constructed from D according to the following rules: Enter -1 in M for each cell entry in D where data exist. Enter 0 in all other off-diagonal cells. The entry in the diagonal cells in n_i—the number of observations in row (or column) i. Note that M is a symmetric matrix and that the sum of each row (and column) is zero, and hence M^{-1} does not exist.

$Z \equiv D1$, a column vector, the sum of the rows of matrix D.

Using this notation, the set of equations (4) may be written

$$MS = Z. \tag{5}$$

Matrix M has no inverse, but, in general, its first minor will have an inverse. If we specify an origin by some method such as arbitrarily setting the first element of S equal to zero, a solution for the remaining elements of S is given by deleting the first element from Z and S, and the first row and column from M, giving

$$M_{11}S_1 = Z_1 , \tag{6}$$

which has the solution

$$S_1 = M_{11}^{-1}Z_1 . \tag{7}$$

It may be noted that for complete data M is an $n \times n$ matrix with $n - 1$ in the principal diagonal and -1 in each off-diagonal cell. The inverse of M_{11} for the complete data case is a matrix with $2/n$ in each principal diagonal cell and $1/n$ in each off-diagonal cell. In this case the solution given in (7) corresponds to the solution given by Mosteller (**5**, equation 10).

Subsequent computations and the present analysis will be facilitated by the use of two other matrices:

$N \equiv$ a diagonal matrix with reciprocal of $(n_i + 1)$ as the element in the ith diagonal cell and zero in each off-diagonal cell.

$L \equiv$ a matrix constructed from M by putting a zero wherever there is a zero in M and $+1$ in all other cells.

Note that

$$M = N^{-1} - L. \tag{8}$$

Using this notation (5) becomes

$$Z = N^{-1}S - LS. \tag{9}$$

For the matrix of complete data L becomes a matrix with unity in every cell and thus LS is a vector of constants. The origin of S can be chosen so that $LS = 0$, whereupon (9) can be solved giving

$$S = NZ, \tag{10}$$

which is also equivalent to the solution given by Mosteller (**5**, equation 10).

Iterative Solution

However, for the case of incomplete data where M_{11}^{-1} is difficult to compute, a solution may be found quickly by an iterative procedure. The procedure given here is an analytical analog of a graphical iterative procedure that was devised by Mrs. Gertrude Diederich. It was suggested by the procedure outlined by Garner and Hake (**1**). It corresponds essentially to taking their solution as a *first approximation* and then correcting it to obtain successively closer approximations.

The iterative procedure proposed here for paired comparison is outlined below. We begin by taking a trial set of scale values. Since this corresponds to *assuming* a set of values for S we may say that T_1 designates this first set of trial values. The discrepancies between the *predicted values* (MT_1) and the *experimental values* (Z) are found by taking differences ($Z - MT_1$) $\equiv ME_1$. The average of the discrepancies for each scale value is then used to correct that value. This corresponds to taking the correction NME_1 and computing the second set of trial values by setting $T_2 = T_1 + NME_1$.

Graphically this correction corresponds to changing each element of T_1 by the average of all the discrepancies found for that element. When tried graphically, such a procedure gave convergence as far as detectable from a reasonably large graph in two or three trials. Although no analytical proof of convergence has yet been found, it seems intuitively reasonable that utilizing the same correction analytically would give a reasonable approximation to the solution. In any case, the discrepancies (ME_i) would be computed at each step, so that failure of convergence could be readily detected. This process of computing $T_{i+1} = T_i + NME_i$ is continued until the discrepancies and the correction terms are negligible.

A little matrix substitution shows that if $T_3 = T_2 + NME_2$, then $T_3 = T_1 + N(I + LN)ME_1$. Thus, since the "two step" correction is very easy to compute, we find that $T_2 = T_1 + N(I + LN)ME_1$. In the computational procedure it is also necessary to adjust the *general scale* (variance) of the trial values to agree with that of the observations. Generalizing these computations we have the following steps for an iterative computation of scale values:

1. Select T_1, , any set of trial S values. The values 0, 1, 2, 3, 4, $\cdots$, $(n - 1)$ will suffice although convergence will probably be slightly more rapid if the average difference of paired D_{ij}-values is used.

2. Compute $MT_{1'}$.
3. Adjust the general scale of $T_{1'}$ to Z by computing

$$a^2 = Z'Z/(T_{1'}'M'MT_{1'}).$$

4. Set $T_1 = aT_{1'}$, and $MT_1 = aMT_{1'}$.
5. Compute the error of approximation

$$ME_1 = Z - MT_1.$$

6. Compute the correction

$$C_1 = N(I + LN)ME_1 = N(I + I - MN)ME_1 .$$

7. $$T_{2'} = T_1 + C_1 + b_1 ,$$

where b is an additive constant to adjust the origin to some convenient value.

Repeat beginning with Step 1 until the error of approximation obtained in Step 5 is negligible.

As a guide to computation it should be noted that C is computed most readily by the following sequence: multiplying to obtain NME_1 ; selective addition of elements in NME_1 yields $LNME_1$; another addition of ME_1 to $LNME_1$ gives $(I + LN)ME_1$; multiplying gives $N(I + LN)ME_1$. In two problems worked by this procedure each element of the vector indicated in Step 5 was reduced to .005 or less on the third approximation. One problem is presented here to illustrate the results of this procedure.

Illustrative Problem

A food preference questionnaire was constructed using five different main courses and the ten composites formed by taking all possible pairs of these five. For each choice of the form (i vs. j) or (i and j vs. g) or (i and j vs. g and h) the subject was asked to indicate his preference. Three sample choices are shown in Table 1 together with the code used in Tables 2, 3, and 4 for each of the five foods.

This questionnaire was given to 92 college students with the result shown in Table 2. The number in each cell indicates number of votes for the item indicated at the beginning of the row when it was paired with the item at the top of the column. For example, when given the choice between Tongue and Pork, 68 persons chose Pork while 24 chose Tongue. Four of the paired comparisons were made by only 91 persons. These were utilized rather than lowering the number of complete cases to 88.

Comparisons of the form (i and j vs. i) were omitted from the questionnaire. As the results turned out, it would have been interesting to have had such items. A few comparisons of the form (i and j vs. i and k) were included. These, however, were eliminated in the analysis since it was not clear from the results whether the subjects were judging one *composite* against another or were merely ignoring the common element and comparing (j vs. k) as a (1 vs. 1) comparison. The item (P and S vs. T and L) was given in the middle

TABLE 1

Sample Questionnaire Items

			Code used in Tables 2, 3, and 4
38.	☐	Roast Rib of Prime Beef	(B)
	☐	Roast Loin of Pork	(P)
39.	☐	Roast Rib of Prime Beef	(B)
		Sirloin Steak	(S)
	☐	Boiled Smoked Beef Tongue	(T)
		Loin Lamb Chop	(L)
40.	☐	Sirloin Steak	(S)
	☐	Roast Loin of Pork	(P)
		Boiled Smoked Beef Tongue	(T)

TABLE 2

Food Preference Data*

	TP	T	TL	P	TB	PL	L	TS	PB	B	PS	LB	S	LS	BS
TP	X	-	-	-	-	-	18	-	-	2	-	5	2	4	0
T	-	X	-	24	-	16	13	-	8	1	9	3	0	1	1
TL	-	-	X	37	-	-	-	-	13	5	10 13	-	2	-	3
P	-	68	54	X	39	-	21	30	-	4	-	3	0	6	5
TB	-	-	-	53	X	37	40	-	-	-	17	-	16	9	-
PL	-	76	-	-	55	X	-	45	-	22	-	-	12	-	2
L	73	79	-	71	52	-	X	46	31	13	22	-	6	-	10
TS	-	-	-	62	-	47	46	X	31	32	-	24	-	-	-
PB	-	84	78	-	-	-	60	61	X	-	-	-	21	15	-
B	90	91	87	88	-	70	79	60	-	X	35	-	19	23	-
PS	-	83	82 79	-	75	-	70	-	-	57	X	43	-	-	-
LB	87	89	-	89	-	-	-	68	-	-	49	X	37	-	-
S	90	92	90	92	76	80	86	-	71	73	-	55	X	-	-
LS	88	91	-	86	83	-	-	-	77	69	-	-	-	X	-
BS	92	91	89	87	-	90	82	-	-	-	-	-	-	-	X

*Number of votes for stimulus listed at side when it was paired with the stimulus listed at the top. The key for stimulus abbreviations is given in Table 1.

of the questionnaire and again near the end with the results shown in Table 2. In converting this cell to a "distance" entry the average of the two different distances was used.

The values shown in Table 2 were converted to normal deviate values by use of a table of the normal curve. An adjustment for different discriminal dispersions was made on the basis of a priori considerations. Since the number of items involved in the three types of judgments [the (1 vs. 1), the (2 vs. 1), and the (2 vs. 2)] were in the ratio of 2 to 3 to 4, the *variances*, or the squares of the discriminal dispersions, might reasonably be assumed to have the same ratio. To approximate this ratio the entries from the normal table were used directly for all comparisons of the (2 vs. 1) and (1 vs. 2) type. For comparisons of the (2 vs. 2) type these values were multipled by 1.2 (which equals approximately $\sqrt{4}/\sqrt{3}$). For comparisons of the (1 vs. 1) type the normal curve values were multiplied by 0.8 (which equals approximately $\sqrt{2}/\sqrt{3}$).

Also, the three types of judgments were scaled separately and the scales seemed to agree reasonably well after the adjustment for the assumed differences in discriminal dispersions were made.

Judgments of the form 92 vs. 0 were converted into "lower bound" values by assigning the normal deviate for 91.5 vs. 0.5. As the results turned out, these values were not systematically lower than predicted values so perhaps such an approximation is possible where only a few 100 per cent judgments are found.

Table 3 gives the resulting D matrix, provided a zero is substituted for each missing entry. The sums of the rows shown in Table 3 are elements of the column vector Z.

The results of the computations indicated in Steps 1 to 7 are shown in Table 4. Since a in Step 3 turned out to be nearly unity each time, only the trial vectors T_1, T_2, and T_3 are shown while $T_{1'}$, $T_{2'}$, and $T_{3'}$ are not shown. The vectors indicating the error of approximation to a least squares fit, ME_1, ME_2, and ME_3 are shown, and indicate extremely rapid convergence.

The correction vectors $C_1 + b_1$ and $C_2 + b_2$ indicated in Step 7 are also shown. It should be noted that each b_i has been chosen so that the element corresponding to "Tongue and Pork" (TP) remains at zero. Some such adjustment facilitates comparisons of the successive approximations and does not affect goodness of fit since the origin is arbitrary.

A very interesting regularity appears in these results. The components in order from least preferred to most preferred are Tongue, Pork, Lamb, Beef, and Steak. Adding Lamb, Beef, or Steak *increases* the value of a composite; while adding Tongue or Pork *decreases* the value of the composite. Thus, the evidence, purely from the general ordering of the stimuli, places Tongue and Pork as negative and Lamb, Beef, and Steak as positive values. A more precise determination of the zero point will be given in a later article.

TABLE 3

Food Preferences

Experimental Inter-Stimulus Distances

(The key for stimulus abbreviations is given in Table 1)

	TP	T	TL	P	TB	PL	L	TS	PB	B	PS	LB	S	LS	BS	Sums
TP	0	X	X	X	X	X	-0.850	X	X	-2.020	X	-1.920	-2.020	-2.052	-3.060	-11.922
T	X	0	X	-0.512	X	-0.940	-0.856	X	-1.360	-1.832	-1.290	-1.840	-2.040	-2.290	-2.290	-15.250
TL	X	X	0	-0.240	X	X	X	X	-1.270	-1.600	-1.380	X	-2.020	X	-2.208	-8.718
P	X	0.512	0.240	0	-0.190	X	-0.592	-0.450	X	-1.368	X	-1.840	-2.040	-1.510	-1.600	-8.838
TB	X	X	X	0.190	0	-0.300	-0.160	X	X	X	-1.080	X	-0.940	-1.548	X	-3.838
PL	X	0.940	X	X	0.300	0	X	-0.036	X	-0.710	X	X	-1.120	X	-2.424	-3.050
L	0.850	0.856	X	0.592	0.160	X	0	0.000	-0.410	-0.856	-0.710	X	-1.208	X	-1.230	-1.956
TS	X	X	X	0.450	X	0.036	0.000	0	-0.504	-0.390	X	-0.768	X	X	X	-1.176
PB	X	1.360	1.270	X	X	X	0.410	0.504	0	X	X	X	-0.740	-1.176	X	1.628
B	2.020	1.832	1.600	1.368	X	0.710	0.856	0.390	X	0	-0.300	X	-0.656	-0.670	X	7.150
PS	X	1.290	1.380	X	1.080	X	0.710	X	X	0.300	0	-0.096	X	X	X	4.664
LB	1.920	1.840	X	1.840	X	X	X	0.768	X	X	0.096	0	-0.250	X	X	6.214
S	2.020	2.040	2.020	2.040	0.940	1.120	1.208	X	0.740	0.656	X	0.250	0	X	X	13.034
LS	2.052	2.290	X	1.510	1.548	X	X	X	1.176	0.670	X	X	X	0	X	9.246
BS	3.060	2.290	2.208	1.600	X	2.424	1.230	X	X	X	X	X	X	X	0	12.812

TABLE 4

Iterative Procedure for Determining Scale Values

	$Z - MT_1$	$N(I + LN)ME_1$	$T_1 + C_1$	$Z - MT_2$	$N(I + LN)ME_2$	$S \cong T_3$	$Z - MT_3$
Stimuli[a]	T_1[b]	ME_1[c]	C_1[d]	T_2[b]	ME_2[c]	C_2[d]	T_3[b]	ME_3[c]
TP	0.0	.778	.00000	.000	.049	.00000	.000	.003
T	0.3	-.350	-.15379	.146	-.010	-.00858	.137	.002
TL	0.4	.082	-.12091	.279	-.004	-.00945	.270	-.004
P	0.7	-.138	-.15236	.548	.020	-.00699	.541	.002
TB	1.0	-.338	-.15929	.841	-.031	-.01126	.830	-.005
FL	1.0	.550	-.06575	.934	.027	-.00594	.928	.002
L	1.2	-.156	-.14974	1.050	.020	-.00681	1.043	.003
TS	1.3	-.476	-.19896	1.101	-.035	-.01331	1.088	-.005
FB	1.6	-.072	-.14123	1.459	-.017	-.01087	1.448	-.001
B	1.9	-.350	-.14431	1.756	-.031	-.00986	1.746	-.002
FS	1.9	.164	-.11220	1.788	.008	-.00760	1.780	.003
LB	2.1	.114	-.10014	2.000	.002	-.00709	1.993	-.001
S	2.3	.234	-.09550	2.205	-.003	-.00793	2.197	.000
LS	2.4	.346	-.07185	2.328	.028	-.00364	2.324	.004
BS	2.8	-.388	-.16790	2.362	-.023	-.00986	2.622	-.001

[a] Stimuli abbreviations are defined in Table 1.

[b] T_1, T_2, and T_3 are successive trial values for S (scale values of the stimuli)

[c] ME_1, ME_2, and ME_3 are the errors of approximation.

[d] C_1 and C_2 are the correction terms.

It turned out to be necessary to carry the computations of Step 6 to five decimals in order to be certain of having the new trial value for T correct to three decimal places. Amount of lowest entry has been subtracted from each element in C so that correction on lowest term is zero. This does not affect fit but merely keeps origin of T_i constant.

REFERENCES

1. Garner, W. R., and Hake, H. W. The amount of information in absolute judgments. *Psychol. Rev.*, **58**, 1951, 446–459.
2. Horst, P. The prediction of personal adjustment. *Social Science Research Council Bulletin No. 48*, 1941.
3. Kempthorne, O. The design and analysis of experiments. New York: Wiley, 1952, pp. 96–98.
4. Kendall, M. G. The advanced theory of statistics, Vol. II. London: Griffin, 1948, pp. 229–230.
5. Mosteller, F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**, 1951, 3–9.
6. Thurstone, L. L. The zero point in scaling. (Unpublished study.)

Manuscript received 11/29/54

Revised manuscript received 4/29/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

PROPORTIONAL PROFILES AND LATENT STRUCTURE

W. A. GIBSON

PERSONNEL RESEARCH BRANCH
DEPARTMENT OF THE ARMY*

The identity of problem and solution in Lazarsfeld's latent structure analysis and Cattell's proportional profiles is pointed out. Anderson's latent structure solution is adapted to proportional profiles to yield a possible solution for the communality and rotational problems in factor analysis. A numerical example of the latter is provided.

A Formal Identity

Two recent articles (**4, 6**) have presented identical solutions to the same geometric problem which has appeared in two different contexts. The two different contexts are Lazarsfeld's new latent structure model (**7**) and Cattell's concept of proportional profiles in factor analysis (**3**). The common problem may be stated as follows:

> Given two Gramian matrices, R_1 and R_2 , of order n and rank m, to find corresponding orthogonal factor matrices, V_1 and V_2 , that are proportional by columns and are known to exist from theoretical considerations.

Let F_1 be an arbitrary orthogonal factorization of R_1 , and let Λ be the orthogonal transformation from F_1 to V_1 . Also let D be the diagonal matrix of proportionality constants. Then

$$R_1 = F_1F_1' = V_1V_1' , \tag{1}$$

$$V_1 = F_1\Lambda, \tag{2}$$

$$V_2 = V_1D = F_1\Lambda D, \tag{3}$$

and

$$R_2 = V_2V_2' = V_1D^2V_1' = F_1\Lambda D^2\Lambda'F_1' . \tag{4}$$

The solution is to form the matrix

$$\begin{aligned} A &= (F_1'F_1)^{-1}F_1'R_2F_1(F_1'F_1)^{-1} \\ &= (F_1'F_1)^{-1}F_1'F_1\Lambda D^2\Lambda'F_1'F_1(F_1'F_1)^{-1} \\ &= \Lambda D^2\Lambda', \end{aligned} \tag{5}$$

*This paper was initiated at the University of North Carolina and completed at the Center for Advanced Study in the Behavioral Sciences.

factor it by any principal components method, and normalize the resulting factor matrix by columns to obtain Λ. Then V_1 and V_2 are given by (2) and (3) in turn, D having been determined in the normalizing process.

Since the diagonal elements in D^2 are the characteristic roots of A, this solution cannot yield a unique Λ unless all diagonal elements in D^2 (hence also in D) are different. This creates no serious difficulties in latent structure analysis, for there the empirical data can usually be combined in such a way as to make all diagonal values in D^2 quite distinct from each other. In proportional profiles, however, there is no such manipulatory freedom. There R_1 and R_2 are correlation matrices for the same tests based on two different samples, and V_1 and V_2 are corresponding factor matrices. The diagonal values in D indicate differential degrees of selection on the factors in samples 1 and 2. In practice it may often be difficult to find two samples in which the degrees of selection on the various factors are all quite different. This will create problems of slow convergence or even indeterminacy for some of the factors. While such problems will not be discussed here, their possible seriousness should not be minimized in evaluating the notion of proportional profiles.

It is instructive to consider the special case where F_1 is the principal components analysis of R_1 . The matrix $F_1'F_1$ then is diagonal and contains the characteristic roots of R_1 in its diagonal cells. The matrix $(F_1'F_1)^{-1}$ is also diagonal, with the reciprocals of the characteristic roots of R_1 as its diagonal entries. The pre- and post-multiplication of the matrix $F_1'R_2F_1$ by this diagonal matrix in (5) produces a situation which is exactly the opposite of what might be expected. The last diagonal values in $(F_1'F_1)^{-1}$ are the largest, so that the last vectors in A are likely to be the longest. Hence the first principal components in A are determined largely by the last principal components in R_1 . Since the last components in R_1 are probably most seriously affected by error, less confidence can be placed in the first columns of V_1 and V_2 than in the last. This line of reasoning applies also to the centroid F_1 insofar as it approximates the principal components F_1 .

Anderson's Latent Structure Solution and Proportional Profiles

Since the geometric problem in latent structure analysis and in proportional profiles is the same, it follows that any solution for the latent structure equations is also a solution for proportional profiles. A recent solution for the former, developed by Anderson (1) and extended by Gibson (5), has the advantage over Green's solution of avoiding the estimation of any unknown elements in the manifest matrices (such as the missing diagonal terms). It is not difficult to adapt this solution to proportional profiles so as to eliminate both the communality and rotational problems in factor analysis, if Cattell's concept is accepted and is workable.

Let the two correlation matrices, R_1 and R_2 , be rearranged, if need be,

so that the two sub-matrices of order $n - m$ by m in the lower left corner of R_1 and R_2 have rank m. Let these two sub-matrices be designated P_1 and P_2, respectively. Let the subscript a refer to the last $n - m$ test variables, and let b stand for the first m test variables. Then

$$P_1 = V_{a1}V'_{b1}, \tag{6}$$

and

$$P_2 = V_{a2}V'_{b2} = V_{a1}D^2V'_{b1}. \tag{7}$$

That is, V_{a1} and V_{a2} are made up of the last $n - m$ rows of V_1 and V_2, respectively, and V'_{b1} and V'_{b2} are square matrices consisting of the first m columns of V'_1 and V'_2, respectively.

Now form the matrix

$$\begin{aligned} B &= (P'_1P_1)^{-1}P'_1P_2 \\ &= (V_{b1}V'_{a1}V_{a1}V'_{b1})^{-1}V_{b1}V'_{a1}V_{a1}D^2V'_{b1} \\ &= V'^{-1}_{b1}(V'_{a1}V_{a1})^{-1}V^{-1}_{b1}V_{b1}V'_{a1}V_{a1}D^2V'_{b1} \\ &= V'^{-1}_{b1}D^2V'_{b1}. \end{aligned} \tag{8}$$

The next task is to obtain the characteristic roots and a set of right-sided characteristic vectors of B, for it turns out that the roots are the diagonal entries in D^2 and the vectors are the columns of $V'^{-1}_{b1}K$, K being an arbitrary unknown diagonal. Thus

$$\begin{aligned} BV'^{-1}_{b1}K &= V'^{-1}_{b1}D^2V'_{b1}V'^{-1}_{b1}K \\ &= V'^{-1}_{b1}D^2K = V'^{-1}_{b1}KD^2. \end{aligned} \tag{9}$$

Once a matrix $V'^{-1}_{b1}K$ is obtained, (6) can be post-multiplied by it to give

$$C = P_1V'^{-1}_{b1}K = V_{a1}V'_{b1}V'^{-1}_{b1}K = V_{a1}K. \tag{10}$$

Thus V_{a1} becomes available except for unknown multipliers on its columns. It happens that these multipliers are quite readily obtained in latent structure analysis because of the way in which the empirically given matrices are bordered. It is necessary to adopt a different approach in proportional profiles. Let the symmetric sub-matrix of order $n - m$ in the lower right corner of R_1 be designated Q_1. Then

$$Q_1 = V_{a1}V'_{a1} = (V_{a1}K)K^{-2}(KV'_{a1}) = CK^{-2}C'. \tag{11}$$

Except for its diagonal terms, the matrix Q_1 is entirely given, as is the matrix C. Assume for the moment that the diagonal terms in Q_1 are given, and define

$$G = C(C'C)^{-1}. \tag{12}$$

Then K^{-2} can be obtained from (11) as follows:

$$K^{-2} = (C'C)^{-1}C'Q_1C(C'C)^{-1} = G'Q_1G. \tag{13}$$

Given K^{-2}, K^{-1} is easily formed, after which V_{a1} can be determined by rewriting (10) in the form

$$V_{a1} = CK^{-1}. \tag{14}$$

This suggests a simple iterative procedure in which V_{a1} is first approximated by inserting rough diagonal estimates into Q_1 and applying (13) and (14), then improved diagonal estimates obtained from the first V_{a1} permit a second cycle of the same kind, and so on until further iteration makes no important change in V_{a1} from one cycle to the next. For large Q_1 matrices, little or no iterating will be needed, since the products involving the estimated diagonals will constitute such a small part of the sums of products. With small Q_1 matrices the iterations will not be very time-consuming. To save time in iterating it may be worth while to express Q_1 as the sum of two matrices, one of them being a diagonal matrix containing the unknown diagonal elements of Q_1, and the other being Q_1 with zero diagonal elements. Designate these two matrices E and Q_{10}, respectively. Then

$$K^{-2} = G'(Q_{10} + E)G = G'Q_{10}G + G'EG. \tag{15}$$

Only the last term in (15) changes from one trial to the next, and it is rapidly formed. The matrix G remains constant throughout the iterative process.

Given V_{a1}, V_{b1} can be solved for from (6), (10), and (12) as follows:

$$\begin{aligned} V_{b1} &= P_1'V_{a1}(V_{a1}'V_{a1})^{-1} \\ &= P_1'V_{a1}K(KV_{a1}'V_{a1}K)^{-1}K \\ &= P_1'C(C'C)^{-1}K = P_1'GK. \end{aligned} \tag{16}$$

The last member of (16) has been adjusted so as to involve the same matrix G that is used in the iteration to determine V_{a1}. Thus G is made to serve two purposes.

Once V_{a1} and V_{b1} are known, all that remains to be done is to form V_1 from them to give the "true" factorization of R_1, and then to compute V_2 from the first part of (3), the matrix D having been formed from the square roots of the characteristic roots of B. The goodness of fit of the two factorizations V_1 and V_2 is indicated by the agreement between the first and third members of (1) and between the first and second members of (4).

It is to be noted that this adaptation of Anderson's solution to the problem of proportional profiles assumes the number of factors to be known at the start. This is not a serious drawback, however, for there are many ways of estimating the rank of a correlation matrix, and the inverting of $P_1'P_1$ can

be done in such a way (cf. **8**, pp. 46-48) that a change in its order can be accomplished without serious loss of computing time.

Several equations in this paper have least squares properties that should perhaps be listed explicitly. The first is (5), which is the best fitting solution for $\Lambda D^2 \Lambda'$ in (4). The next is (8), which is the least squares solution of B in the equation $P_2 = P_1 B$. Then (13) is the best fitting solution for K^{-2} in (11). Finally, (16) is the least squares solution for V_{b1} in (6).

A Fictitious Example

As an example of the present solution for proportional profiles, consider the two fictitious correlation matrices, R_1 and R_2, that are shown in Tables 1 and 2. They were generated from a simple structure V_1 and V_2 that originally were strictly proportional by columns, but that subsequently were "blurred" by adding small random increments to each of their entries. Thus there exists no perfect proportional profiles fit for R_1 and R_2, so that the various least squares properties of the solution will have definite advantages in this example.

Inspection of R_1 and R_2 suggests that three factors will probably account for both, and that the 7×3 sub-matrix in the lower left corner of each probably has rank 3. These two sub-matrices are therefore designated P_1 and P_2, respectively, and (8) is used to form from them the matrix B that is shown in Table 3.

Table 4 is the matrix D^2, containing in its diagonal cells the characteristic roots of B. These characteristic roots are obtained by forming and solving the characteristic equation of B(cf. **8**, pp. 26-28 and 44-45). The algebraic sign of the characteristic roots in Thurstone's discussion is the opposite of the convention being used here. The three columns of Table 5 are a set of right-sided characteristic vectors of B. Column A, for example, is a solution to the set of homogeneous linear equations whose coefficients matrix is the matrix B with its first characteristic root subtracted from each of its diagonal cells (cf. **2**, pp. 250-251). Column B is the same thing using the second characteristic root, etc. Each such set of homogeneous linear equations is readily converted into a consistent set of m non-homogeneous linear equations in $m - 1$ unknowns by arbitrarily fixing one of the unknowns at some convenient value. Here the first element of the first characteristic vector was set equal to unity. The resulting over-determined system can be solved by any method. Here it was done by least squares because of the slight inconsistency resulting from rounding.

The next step is to form the matrix C from P_1 and $V_{b1}'^{-1}K$ by means of (10). C is shown in Table 6; (10) indicates that it is proportional by columns to V_{a1}.

The task now becomes one of extracting V_{a1} from C by determining K^{-1}. As a first step the matrix G is formed by means of (12); G is shown in

TABLE 1
A Fictitious R_1

	1	2	3	4	5	6	7	8	9	10
1		.05	.02	.25	.02	.02	.18	.13	.06	.18
2	.05		.10	.07	.46	.08	.36	.11	.24	.31
3	.02	.10		.00	.14	.39	.10	.35	.33	.20
4	.25	.07	.00		.05	.00	.19	.11	.05	.18
5	.02	.46	.14	.05		.12	.39	.14	.28	.34
6	.02	.08	.39	.00	.12		.08	.38	.36	.20
7	.18	.36	.10	.19	.39	.08		.17	.23	.35
8	.13	.11	.35	.11	.14	.38	.17		.35	.27
9	.06	.24	.33	.05	.28	.36	.23	.35		.30
10	.18	.31	.20	.18	.34	.20	.35	.27	.30	
		P_1					Q_1			

TABLE 2
A Fictitious R_2

	1	2	3	4	5	6	7	8	9	10
1		.13	.08	.46	.08	.06	.32	.23	.13	.37
2	.13		.05	.16	.50	.01	.45	.10	.25	.37
3	.08	.05		.08	.10	.20	.07	.24	.20	.15
4	.46	.16	.08		.10	.05	.37	.26	.14	.42
5	.08	.50	.10	.10		.05	.40	.12	.27	.34
6	.06	.01	.20	.05	.05		.02	.20	.16	.10
7	.32	.45	.07	.37	.40	.02		.19	.24	.47
8	.23	.10	.24	.26	.12	.20	.19		.23	.28
9	.13	.25	.20	.14	.27	.16	.24	.23		.28
10	.37	.37	.15	.42	.34	.10	.47	.28	.28	
		P_2								

TABLE 3
$B=(P_1'P_1)^{-1}P_1'P_2$

	1	2	3
1	1.76	.29	.25
2	.07	1.15	.00
3	.04	.18	.55

TABLE 4
D^2

	A	B	C
A	1.80		
B		1.13	
C			.54

TABLE 5
$V_{b1}'^{-1}K$

	A	B	C
1	1.00	-.33	-.21
2	.11	1.00	.02
3	.02	-.33	1.00

TABLE 6
$C=P_1V_{b1}'^{-1}K$

	A	B	C
4	.26	-.01	-.05
5	.07	.41	.14
6	.04	-.06	.39
7	.22	.27	.07
8	.15	-.05	.32
9	.09	.11	.32
10	.22	.18	.17

TABLE 7
$G=C(C'C)^{-1}$

	A	B	C
4	2.35	-.84	-.72
5	-.97	1.76	.24
6	-.32	-.47	1.17
7	.89	.63	-.29
8	.74	-.75	.71
9	-.26	.21	.82
10	.95	.18	.05

Table 7. Now the constant part ($G'Q_{10}G$) of (15) can be formed; it is shown in Table 8.

The initial diagonal estimates for Q_1 might as well be the same first approximations that have served so well with centroid factoring—the highest non-diagonal entry in each of the columns. These need not be taken from

within Q_1 . They can come from the first three rows of R_1 , as they do for columns 4, 5, and 6. The resulting variable part $(G'E_1G)$ of (15) appears in Table 9. The subscript on E_1 indicates that it is the first in a series of E matrices.

The sum of the matrices in Tables 8 and 9 is given by Table 10. The non-vanishing side entries in Table 10 call attention to the fact that there is

TABLE 8
$G'Q_{10}G$

	A	B	C
A	.92	1.13	.83
B	1.13	.37	.24
C	.83	.24	1.61

TABLE 9
$G'E_1G$

	A	B	C
A	2.72	-1.17	-.63
B	-1.17	2.08	-.07
C	-.63	-.07	1.16

TABLE 10
$K_1^{-2}=G'Q_{10}G+G'E_1G$

	A	B	C
A	3.64	-.04	.20
B	-.04	2.45	.17
C	.20	.17	2.77

nothing in the equations which prevents this estimated K^{-2} from being non-diagonal when the data are imperfect. Some reduction of the side entries in K^{-2} may occur as the iterations proceed, but they may never vanish identically.

A first estimate of K^{-1} is obtained by taking the square roots of the diagonal entries in Table 10. Thus the side entries are not used. This suggests that a further shortening of the iterative procedure would be to compute only the diagonal terms in (15).

The first estimate of V_{a1} , obtained by (14), is shown in Table 11. The last column of Table 11 gives the new diagonal estimates (the sums of squared row entries in this first V_{a1}) which make up the E_2 of the second trial. Now the matrices $G'E_2G$, K_2^{-2}, K_2^{-1}, and CK_2^{-1} can be formed in turn exactly as in the first trial, the resulting second estimate of V_{a1} providing an E_3 , etc. In the present example this process was continued for a total of

TABLE 11
First $V_{a1}=CK_1^{-1}$

	A	B	C	$h_1^2=$
4	.50	-.02	-.08	.26
5	.13	.64	.23	.48
6	.08	-.09	.65	.44
7	.42	.42	.12	.37
8	.29	-.08	.53	.37
9	.17	.17	.53	.34
10	.42	.28	.28	.33

TABLE 12
Final $V_{a1}=CK_4^{-1}$

	A	B	C	$h_4^2=$
4	.50	-.02	-.08	.26
5	.14	.66	.24	.51
6	.08	-.10	.66	.45
7	.42	.44	.12	.38
8	.29	-.08	.54	.38
9	.17	.18	.54	.35
10	.42	.29	.29	.35

four trials, at the end of which E_5 became identical with E_4 at the level of accuracy being used, so that no further change in the estimated V_{a1} could take place. This final V_{a1} is shown in Table 12, along with the final diagonal estimates for Q_1 . It will be seen that no essential change has occurred from the first to the final V_{a1} , although all diagonal estimates but the first have

changed. Most of the changes occurred between the first and second estimates of V_{a1} . All changes after the second V_{a1} (six loadings and two diagonals) were of size .01. This suggests that one repetition of the iterative procedure may often suffice for practical purposes.

Now V_{b1} is formed by means of (16) and is recorded in Table 13. Tables 12 and 13 together make up V_1 , so that V_2 can now be obtained by (3). V_2 appears in Table 14. The structure of V_1 and V_2 is further removed than was anticipated from the simple structure configuration which generated this

TABLE 13
$V_{b1}=P_1'GK_4$

	A	B	C
1	.50	-.07	-.03
2	.17	.59	.20
3	.10	-.02	.59

TABLE 14
$V_2=V_1'D$

	A	B	C
1	.67	-.07	-.02
2	.23	.63	.15
3	.13	-.02	.43
4	.67	-.02	-.06
5	.19	.70	.18
6	.11	-.11	.49
7	.56	.47	.09
8	.39	-.09	.40
9	.23	.19	.40
10	.56	.31	.21

example. Apparently the result is highly sensitive to such minor distortions as rounding error and the random increments that were mentioned earlier. Such small changes seem able to force a sizeable shift in the position of the reference frame before the maximum degree of column proportionality is restored. This apparent instability in rotation should perhaps be kept in mind in the application of any proportional profiles or latent structure solution.

Further rotation of V_1 and V_2 would improve their simple structure appearance, but that would of course reduce the extent of their column proportionality. This illustrates the fact that nothing in the equations for proportional profiles guarantees a simple structure solution. The proportional profiles result might be nearer to a centroid or principal components analysis. However, it would seem that if the notion of proportional profiles is to be generally applicable and psychologically meaningful, then the required differential selection should take place with respect to something like the factors of simple structure, where factorial composition is at least partially independent of the make-up of the particular test battery.

The degree of fit for the present example is indicated by the two residual matrices shown in Tables 15 and 16. The difference in goodness of fit for R_1 and for R_2 calls attention to a consideration which is purely empirical. With perfect data it would make no difference which of the two correlation matrices was designed R_1 and which was called R_2 . With empirical data, on the other

hand, such matters as sampling and sample size may appropriately influence the choice. It is to be expected that R_1 will be fitted best, since P_1 and Q_1 are involved in the solutions for C and K^{-1}. It is also to be expected that the P_2 section of R_2 will be fitted better than the rest of R_2, since P_2 is the only part of R_2 that is used in the solution. If it were considered to be worth the trouble, the fit of the two correlation matrices could be more nearly equalized by post-multiplying C by D to form $V_{a2}K$. Then the iterative procedure

TABLE 15
R_1-V_1V_1'

	1	2	3	4	5	6	7	8	9	10
1		.01	-.01	.00	.00	-.01	.00	.00	.00	.00
2	.01		-.02	.01	.00	-.01	.00	.00	.00	.01
3	-.01	-.02		.00	.00	-.01	.00	.00	.00	-.01
4	.00	.01	.00		.01	.01	.00	.01	.01	.00
5	.00	.00	.00	.01		.02	.01	.02	.01	.02
6	-.01	-.01	-.01	.01	.02		.01	-.01	.01	.00
7	.00	.00	.00	.00	.01	.01		.02	.01	.01
8	.00	.00	.00	.01	.02	-.01	.02		.02	.01
9	.00	.00	.00	.01	.01	.01	.01	.02		.02
10	.00	.01	-.01	.00	.02	.00	.01	.01	.02	

TABLE 16
R_2-V_2V_2'

	1	2	3	4	5	6	7	8	9	10
1		.02	.00	.01	.01	-.01	-.02	-.03	.00	.02
2	.02		-.03	.03	-.01	-.02	.01	.01	.02	.01
3	.00	-.03		.02	.01	-.03	-.03	.02	.00	-.01
4	.01	.03	.02		.00	.00	.01	.02	.01	.06
5	.01	-.01	.01	.00		.02	-.05	.04	.02	-.02
6	-.01	-.02	-.03	.00	.02		-.03	-.05	-.04	-.03
7	-.02	.01	-.03	.01	-.05	-.03		-.02	-.01	-.01
8	-.03	.01	.02	.02	.04	-.05	-.02		.00	.01
9	.00	.02	.00	.01	.02	-.04	-.01	.00		.01
10	.02	.01	-.01	.06	-.02	-.03	-.01	.01	.01	

could be applied to obtain a second estimate of K^{-1} based on the Q_2 section of R_2. Some sort of average (possibly the geometric mean) of the two K^{-1} estimates should then lead to a solution fitting the two correlation matrices to about the same extent. The Cattell solution for proportional profiles probably provides better over-all fit, but then R_1 must be factored and the communality problem reappears.

It may be mentioned, in passing, that the usual summational checks are applicable at nearly every stage of the present solution. They have been omitted here to simplify the exposition.

REFERENCES

1. Anderson, T. W. On estimation of parameters in latent structure analysis, *Psychometrika*, 1954, **19**, 1–10.
2. Birkhoff, G., and MacLane, S. A survey of modern algebra. New York: Macmillan, 1953.
3. Cattell, R. B. "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. *Psychometrika*, 1944, **9**, 267–283.
4. Cattell, R. B., and Cattell, A. K. S. Factor rotation for proportional profiles: analytical solution and an example. *Brit. J. statist. Psychol.*, 1955, **8**, 83–92.
5. Gibson, W. A. An extension of Anderson's solution for the latent structure equations. *Psychometrika*, 1955, **20**, 69–73.
6. Green, B. F. A general solution for the latent class model of latent structure analysis. *Psychometrika*, 1951, **16**, 151–166.
7. Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. Chapter 10 in Stouffer, S. A. et al.: Measurement and prediction. Princeton: Princeton Univ. Press, 1950.
8. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. Chicago Press, 1947.

Manuscript received 5/4/54

Revised manuscript received 3/10/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

A NON-PARAMETRIC TEST OF CORRELATION USING RANK ORDERS WITHIN SUBGROUPS

WARREN S. TORGERSON

EDUCATIONAL TESTING SERVICE

Kendall's rank order test for association between two variables is generalized to the case where the total sample is made up of several subgroups and the data on one or both variables consist of the rank order within each subgroup. The test involves no assumptions concerning scales of measurement, shapes of distributions, or relative level of excellence or amount of variability of the different subgroups. Two empirical examples indicate that the normal approximation to the exact test of significance can be considered adequate for most practical situations. Special consideration is given to the case of tied ranks. If ties occur in but one variable within any given subgroup, only a slight modification in procedure is needed. Extensive ties in both variables within subgroups lead to difficulties in determining the appropriate correction for continuity.

I. *Problem*

The problem occasionally arises of determining whether or not a significant correlation exists between two variables when the total sample is made up of a number of subsamples, and the scores on one or both variables consist of rank orders within the subsamples. Such a situation frequently occurs, for example, when one wishes to determine whether or not scores on a psychological test are significantly related to supervisors' rankings with respect to proficiency on some task. Similar situations commonly occur in the military establishment. In the measurement of personality traits it is also desired at times to determine the significance of correlation between a trait measured by a personality schedule or test and ratings or rankings on the trait as made by peers. In any of these situations it frequently happens that the population is best considered as composed of many subgroups, each headed by its own supervisor or officer, or composed of subgroups whose members are well acquainted with one another.

Let us consider the first situation in somewhat more detail. Assume that a total sample is composed of k small subgroups (of from, say, three to six or seven subjects each), each headed by a supervisor. Each supervisor can place his own subordinates in rank order with respect to proficiency on the task of interest. No supervisor, however, has any information concerning proficiencies of subjects supervised by others. Hence, no information is available concerning the relative positions of subjects from different subgroups. It is not known, for example, whether the highest ranked person in one

subgroup is better or worse on the criterion than the lowest ranked person in another subgroup.

We thus have as raw data, first, scores for all subjects on the psychological test, and, second, criterion data in the form of k sets of ranks, one set for each of the k subgroups composing the total sample. The problem is to determine whether or not a significant correlation obtains between the test and the criterion.

When one is confronted with this problem, perhaps his usual procedure is either to decide not to carry out the experiment after all or to use procedures which require one or another of several sets of simplifying assumptions.

One such procedure, for example, would be to throw out enough observations so that each subgroup contains the same number of cases, treat the ranks as if they were a single rank order with k ties at each rank, and then test the significance of the rank correlation. Rather than eliminating cases to equate the size of subgroups, linear transformations could be made on the original ranks in such a way as to equate the average rank and the spread of the ranks for the different subgroups, and the separate ranks then combined into a single rank order. Either of these procedures involves a rather imposing set of assumptions.

A different approach would be to compute a test of significance for each subgroup separately and then combine the tests. Either a binomial test or a chi-square test using the $-2 \log p$ transformation might be used (1). However, due to the extremely small numbers of cases within the separate subgroups, neither would seem especially appropriate here: the binomial test because it uses only a portion of the data, and the chi-square test because the highly discrete nature of the data violates the assumption of a rectangular continuous distribution of probabilities.

The purpose of this note is to present a simple over-all non-parametric test for correlation between the two variables. No assumptions whatsoever are made concerning scales of measurement, shapes of distributions, or relative level of excellence and/or amount of variability between subgroups.

II. *The Test of Significance*

The proposed test is an extension of Kendall's test of association between two rank orders (2). His test involves, first, computation of the Kendall sum and, second, the determination of the probability of obtaining a sum that large or larger by chance alone. If the individuals are placed in rank order with respect to one of the variables, the Kendall sum is the number of pairs of individuals in the same order on the second variable minus the number of pairs of individuals in the reverse order. (It should be noted that special rules for computation make actual counting of all pairs unnecessary.) Since there are $n(n-1)/2$ possible pairs for n subjects, the possible values of the sum range from $-n(n-1)/2$ through $+n(n-1)/2$. The distribution of

possible values of the sum under the null hypothesis of zero correlation between the two variables is symmetrical about zero and has a variance of $n(n - 1)(2n + 5)/18$. Further, the distribution approaches normality very rapidly, so that with 10 or more subjects, the normal approximation gives adequate results. Exact distributions have been computed for n's up to 10 (**2**, p. 141).

The Kendall sum, of course, deals with rank orders of two variables for a single group. What then of our problem concerning several subgroups? Here we find that the Kendall sum has some exceedingly pleasing properties.

Let s_j denote the Kendall sum of the jth subgroup, where $j = 1, 2, \cdots, k$. We can obtain a *sum* of Kendall sums for the k subgroups. Let S denote this sum:

$$S = \sum_{j=1}^{k} s_j \, . \tag{1}$$

The distribution of S under the null hypothesis of zero correlation within each subgroup is also symmetrical about zero and has a variance equal to the sum of the variances of the sums (σ_j^2) of the separate subgroups,

$$\sigma_S^2 = \sum_{j=1}^{k} \sigma_j^2 = \sum_{j=1}^{k} n_j(n_j - 1)(2n_j + 5)/18. \tag{2}$$

Further, the distribution also appears to approach normality very rapidly. Using the normal approximation, the test is simply a test of the significance of the critical ratio

$$\text{CR} = \frac{S \mp 1}{\sigma_S} = \frac{\sum_{j=1}^{k} s_j \mp 1}{\sqrt{\frac{1}{18} \sum_{j=1}^{k} n_j(n_j - 1)(2n_j + 5)}}. \tag{3}$$

The "one," which is subtracted from S whenever S is positive and added to S whenever S is negative, is a correction for continuity. Since possible values of S in any given situation are either all odd numbers or all even numbers with a class interval of two, the correction for continuity, taken as one half the class interval, is unity.

To get an idea of whether or not the distribution approaches normality rapidly enough to be of any use, two empirical tests were run—one using subgroups of sizes 3, 4, 4, and 6, and the other using subgroups of sizes 4, 4, 5, and 5. For every possible value of S the exact probability of obtaining a value of S that large or larger was computed, along with the corresponding probability using the normal approximation. The results are indicated in Table 1. As can be seen, the approximation is excellent even with samples as small as these. Since most practical situations will involve either more subsamples or subsamples of greater size, the normal approximation can be considered adequate.

TABLE 1

Comparison of Probabilities Computed Using the Normal Approximation (P_N) with the Corresponding Exact Values (P_E)*

Example 1			Example 2		
n_j = 3, 4, 4, and 6			n_j = 4, 4, 5, and 5		
S	P_N	P_E	S	P_N	P_E
			32	.0000	.0000
30	.0000	.0000	30	.0000	.0000
28	.0001	.0000	28	.0001	.0000
26	.0002	.0000	26	.0002	.0001
24	.0005	.0002	24	.0006	.0003
22	.0014	.0008	22	.0016	.0010
20	.0034	.0024	20	.0038	.0029
18	.0078	.0064	18	.0085	.0073
16	.0164	.0149	16	.0175	.0164
14	.0321	.0311	14	.0339	.0331
12	.0587	.0588	12	.0612	.0612
10	.1001	.1016	10	.1031	.1042
8	.1594	.1623	8	.1628	.1649
6	.2383	.2417	6	.2414	.2438
4	.3347	.3374	4	.3368	.3388
2	.4435	.4445	2	.4443	.4449

*Only the positive half of the distribution is tabulated, the negative half being obtained by symmetry. P refers to the probability of obtaining a value of S that large or larger under the null hypothesis.

III. *Numerical Example*

We shall apply the test to the set of hypothetical data given in Table 2. The total sample is made up of five subgroups ($j = 1, 2, \cdots, 5$) of 3, 5, 6, 8, and 8 subjects respectively. For each subject in each subgroup, two scores are available, the score on variable 1 being expressed in terms of the rank position of the subject within his own subgroup and the score on variable 2

TABLE 2

Raw Data for Numerical Example*

Group 1		Group 2		Group 3		Group 4		Group 5	
V_1	V_2	V_1	V_2	V_1	V_2	V_1	V_2	V_1	V_2
1	75	1	46	1	60	1	67	1	63
2	62	2	54	2	77	2	38	2	70
3	56	3	65	3	51	3	41	3	68
		4	40	4	47	4	44	4	59
		5	32	5	56	5	43	5	48
				6	49	6	37	6	67
						7	59	7	51
						8	30	8	54

*Scores on variable 1 (V_1) are expressed in terms of rank order within subgroup and on variable 2 (V_2) in terms of test scores.

expressed in terms of ordinary test scores. The subjects have been arranged in order of their rank position within their own subgroup.

The first step is to transform the test scores into ranks within subgroups. This has been done in Table 3. We now compute the Kendall sum for each subgroup. Since the subjects in each subgroup have been ordered with respect to variable 1, this is accomplished most easily by first determining for each subject in the subgroup the number of subjects below him who have a higher rank number on variable 2. For example, in subgroup 2 there are two subjects with higher rank numbers below the first subject, two below the second subject, two below the third subject, one below the fourth, and none below the fifth. If the sum of these numbers is denoted by r, then

$$s = 2r - \tfrac{1}{2}n(n-1).$$

For subgroup 2,

$$r = 2 + 2 + 2 + 1 = 7,$$

and, hence,

$$s_2 = 14 - \tfrac{1}{2}(5)(4) = 4.$$

In like manner, $s_1 = 3$, $s_3 = 7$, $s_4 = 8$, and $s_5 = 12$, and the sum of Kendall sums, $S = 3 + 4 + 7 + 8 + 12 = 34$.

TABLE 3

Computation of Values of s_j and $\underline{S}$*

Group 1		Group 2		Group 3		Group 4		Group 5	
V_1	V_2	V_1	V_2	V_1	V_2	V_1	V_2	V_1	V_2
1	1	1	3	1	2	1	1	1	4
2	2	2	2	2	1	2	6	2	1
3	3	3	1	3	4	3	5	3	2
		4	4	4	6	4	3	4	5
		5	5	5	3	5	4	5	8
				6	5	6	7	6	3
						7	2	7	7
						8	8	8	6

$r_1 = 2+1 = 3$	$s_1 = 6 - 3 = 3$
$r_2 = 2+2+2+1 = 7$	$s_2 = 14 - 10 = 4$
$r_3 = 4+4+2+0+1 = 11$	$s_3 = 22 - 15 = 7$
$r_4 = 7+2+2+3+2+1+1 = 18$	$s_4 = 36 - 28 = 8$
$r_5 = 4+6+5+3+0+2+0 = 20$	$s_5 = 40 - 28 = 12$
	$\underline{S} = \overline{34}$

*Scores on both variables expressed in terms of rank order within subgroup.

The variance of S under the null hypothesis is equal to the sum of the variances of the five separate subgroups. From equation 2 we have

$$\sigma_S^2 = \tfrac{1}{18}(66 + 300 + 510 + 1176 + 1176) = 179.3333.$$

From (3),

$$\mathrm{CR} = \frac{34 - 1}{\sqrt{179.3333}} = 2.464$$

which, using a two-tailed test, is significant beyond the .02 level.

IV. *Tied Ranks*

Heretofore we have assumed implicitly that ties would not occur. When ties do occur several alternative procedures are available. One such procedure would be to assign ranks within ties randomly and then proceed as if no ties had occurred. A second procedure might be to assign ranks within ties in the most favorable way and also in the least favorable way and then run a separate test of significance on each. If both tests lead to the same conclusion all is well. However, this procedure is inadequate if the two tests lead to conflicting decisions.

A third procedure is to assign the mean rank to the ties. This is the procedure adopted by Kendall for the single group case.

The mean rank method, or mid-rank method as it has been called, requires some additional considerations in computation of both the Kendall sum and its variance for the individual subgroup. Two cases need to be distinguished: (*i*) that in which ties occur, for any given subgroup, in only one of the two variables; (*ii*) that in which ties occur in both variables within a subgroup.

(*i*) *Ties in one variable only for any given subgroup*

It should be noted at the onset that it is irrelevant whether the variable containing the ties is the same variable in all subgroups, or whether for some subgroups it is one variable and for others, the other variable.

The Kendall sum, s_j for any subgroup j has the same meaning as before with the additional stipulation that tied pairs, being neither in the same nor in the reverse order on the second variable, are counted as zero. If the subjects are arranged in order on the variable containing no ties, the sum is the number of pairs on the second variable that are in the same order minus the number of pairs in the reverse order. Tied pairs neither add to nor subtract from the sum.

Calculation of the variance of the jth subgroup requires the following modification. There may be several different ties, and any number of subjects may be involved in any given tie. Let a be an index referring to a particular tie ($a = 1, 2, \cdots, g$) and t_{aj} refer to the number of subjects involved in the ath tie of the jth subgroup. Then Kendall has shown that

$$\sigma_j^2 = \frac{1}{18}\left[n_j(n_j - 1)(2n_j + 5) - \sum_{a=1}^{g} t_{aj}(t_{aj} - 1)(2t_{aj} + 5)\right]. \quad (4)$$

The sum of Kendall sums and its variance under the null hypothesis remain the sum of the individual sums and the sum of the individual variances, respectively. The distribution remains symmetrical about zero and also appears to approach normality rapidly. Possible values of S for any particular combination of number and size of subgroups remain either all odd or all even, with a step interval of two. Hence for the normal approximation, the

correction for continuity remains one, and the critical ratio,

$$\text{CR} = \frac{\sum_{i=1}^{k} s_i \mp 1}{\sqrt{\sum_{i=1}^{k} \sigma_i^2}}, \tag{5}$$

where the individual subgroup sums and variances are computed as noted above.

(ii) Ties in both variables for some or all subgroups

The computation of the separate Kendall sums is the same as before except that ties in *either* rank do not contribute to the sum (see **2**, p. 26, for more detail). The formula for the separate variances is given below, where t_{ai} refers to one variable and u_{bi} to the other.

$$\begin{aligned}\sigma_i^2 = {}& \tfrac{1}{18}[n_i(n_i - 1)(2n_i + 5) \\ & - \sum_a t_{ai}(t_{ai} - 1)(2t_{ai} + 5) - \sum_b u_{bi}(u_{bi} - 1)(2u_{bi} + 5)] \\ & + \frac{1}{9n_i(n_i - 1)(n_i - 2)} [\sum_a t_{ai}(t_{ai} - 1)(t_{ai} - 2)] \\ & \qquad \cdot [\sum_b u_{bi}(u_{bi} - 1)(u_{bi} - 2)] \\ & + \frac{1}{2n_i(n_i - 1)} [\sum_a t_{ai}(t_{ai} - 1)][\sum_b u_{bi}(u_{bi} - 1)].\end{aligned} \tag{6}$$

Again S is the sum of the separate sums, and its variance the sum of the separate variances. The distribution is again symmetrical about zero. However, the correction for continuity depends upon the particular situation. While in the extreme case (a single subgroup with both variables dichotomized) the appropriate correction gets as high as $n/2$, it may be surmised that, for practical situations where there are numerous subgroups and but very few ties occurring in both ranks within the same subgroup, the unit correction for continuity will not be too bad an approximation. An alternative procedure would be to convert the problem to the "ties in one variable only for any subgroup" case by assigning a rank order randomly to the ties in one variable.

REFERENCES

1. Jones, L. V., and Fiske, D. W. Models for testing the significance of combined results. *Psychol. Bull.*, 1953, 50, 375–382.
2. Kendall, M. G. Rank correlation methods. London: Griffin, 1948.

Manuscript received 6/6/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

MAXIMIZING TEST VALIDITY BY ITEM SELECTION

HAROLD WEBSTER

MELLON FOUNDATION

VASSAR COLLEGE

The exact condition for discarding k items from a test in order to obtain a residual test with higher validity is derived. A proof that validity always increases is given for the case $k = 1$. The lack of uniqueness of maximum validity when achieved by use of the condition is discussed. With the use of additional restrictions on items to be included in the initial test, a practical test construction procedure which has several advantages over previous methods is developed. The homogeneity of tests constructed by the method is discussed, and applications are given.

The problem of selecting from among a large number of test items those yielding a test score which will correlate maximally with an external variable has been studied by a number of writers. Gleser and DuBois (6) and Gulliksen (8) have summarized pertinent research. The problem is of first importance when it is necessary to construct a test which has a high correlation with some known criterion; the same problem would arise, however, if it were necessary to select those parts of a composite experimental criterion which would correlate highest with a fixed test. Mathematical solutions to either problem, or to the simultaneous solution of both problems (10), are known. These solutions are impractical in the case of item selection, however, because of the large number of partial regression weights required for items.

An interesting method of testing composites for sampling stability is known (3) but its application for comparing numerous item composites would present serious difficulties. Recently Lord (12) has derived sampling variances for test statistics under conditions of item sampling, but presumably the behavior of the validity coefficient when there is sampling of both persons and items is still unknown. Guilford (7) advises that the approximate solutions be applied only with large tests which have been administered to large numbers of subjects. Richardson and Adkins (13) suggest that possibly any item selection index would be so susceptible to sampling fluctuations that a choice among methods of selecting items would be practically a matter of no importance. Tukey (15) includes item selection for validity among the important unsolved problems of experimental statistics.

Because of difficulties inherent in exact solutions, and despite lack of precise knowledge of sampling fluctuations of item composites, investigators

have devised various approximation methods. Toops (**14**) advocated the construction of tests starting with one highest correlating item and adding, one at a time, items which most increased the validity of the composite. Adkins has applied this method and she and Richardson (**13**) found that a much less laborious modification also gave good results. They selected from a test a number of items at once with highest criterion partial regression weights, the test comprising the other independent variable. The approximate item weight used requires more computing than the one derived in the present paper. Most of the earlier methods, including Toops', assumed a unique order for the item selection; for example, selected items were not reexamined at a later stage to see if they still belonged in the regression structure. Because of random fluctuation in subsequent samples what is gained by such reexamination may indeed not be worth the additional labor; but if the method of reexamination were brief enough, it would seem worth while to use it.

Horst (**9**) advises starting with a relatively large test and rejecting those items for which the ratio of a function of item-criterion covariance to a function of item-test covariance is small. The items are plotted and then selected using the functions as coordinates. Horst cautions against discarding too many items at once without recalculating item-test parameters. Gulliksen (**8**) has simplified Horst's procedure somewhat, but it would still seem better to avoid plotting if a more economical method of identifying items to be rejected can be found.

Flanagan (**4**) constructed tests by retaining those items for which the item-criterion correlation exceeded the item-test correlation; he advised repeating the process, but did not indicate whether he thought it advisable to reexamine previously rejected items to see whether they should be put back into the test. Gleser and DuBois (**6**) recommend what is essentially Flanagan's procedure for a first approximation, and a more refined one for subsequent iterations. It is unclear how they would treat items with positive validities and negative test correlations; rejection of such items always decreases test validity. Their suggestion that the initial test be restricted to positively valid items seems to be a good one, but selection conditions were not derived under this restriction. Gleser and DuBois also state some conditions under which item variances may be ignored in the first approximation to the final test; with the method of the present paper item variances are not required.

In the present paper the exact item selection condition is derived, and the approximations which are used can be seen to be close enough and at the same time to require less computation in applications than any previous ones. Also, a study of the selection condition itself reveals why the traditional expedient of imposing a validity condition on *individual* items to be admitted into a test is reasonable.

I. *Derivation and Properties of the Exact Condition for Discarding k Items in Order to Obtain a Residual Test of Higher Validity*

In order to construct a test with high validity, first a large number n of potential test items is examined for correlation with the external criterion C. It is then possible, with no loss of generality, to score each item j so that its covariance C_{jC} with the criterion C is positive or zero. The sum of these n item scores, $x_1 + \cdots + x_i + \cdots + x_n$, forms an experimental test T from which it is desired to discard items such that the residual test M correlates maximally with C. The correlation of T with C may be written

$$r_{TC} = \sum_{i}^{n} C_{iC}/S_T S_C = C_{TC}/S_T S_C , \tag{1}$$

where C's are covariances, S's standard deviations, and the summation is over the n item scores $1, \cdots, i, \cdots, n$. If the first k items, designated $1, \cdots, j, \cdots, k$, are discarded from T, the validity of the residual test M is

$$r_{MC} = (C_{TC} - \sum_{i}^{k} C_{iC})/S_M S_C . \tag{2}$$

Since M is to be more valid than T, the condition

$$r_{TC} < r_{MC} \tag{3}$$

must be satisfied. Using (1) and (2), (3) may be written

$$C_{TC}/S_T S_C < (C_{TC} - \sum^{k} C_{iC})/S_M S_C . \tag{4}$$

Making the substitution $S_{T-k} = S_M$, and multiplying by S_C, (4) becomes

$$C_{TC}/S_T < (C_{TC} - \sum^{k} C_{iC})/S_{T-k} . \tag{5}$$

As a consequence of having scored items so that $C_{iC} \geq 0$, the quantity in parentheses in (5) must, like the other terms, be positive. The terms of (5) can be rearranged as follows,

$$S_{T-k}/S_T < 1 - [(\sum^{k} C_{iC})/C_{TC}],$$

$$(\sum^{k} C_{iC})/C_{TC} < 1 - [S_{T-k}/S_T],$$

$$S_T/C_{TC} < (S_T - S_{T-k})/\sum^{k} C_{iC} . \tag{6}$$

Writing (6) in a form which will be more useful later,

$$C_{TC}/S_T > (\sum^{k} C_{iC})/(S_T - S_{T-k}). \tag{7}$$

The practical difficulty encountered in using (7) as a condition to be satisfied by rejecting k items in order to enhance validity lies in the fact that

the expansion of S_{T-k} contains the sum of inter-item covariances of the k items. Use of inter-item relationships is laborious and seldom feasible in the typical test construction situation where there is available at most a machine for counting item responses. Further study of (7) is also warranted because it is not clear whether any method exists for identifying uniquely a set of k items to be rejected.

Squaring (5), substituting $V_{T-k} = V_T + V_k - 2C_{kT}$, and simplifying gives

$$C_{TC}^2(V_k - 2C_{kT}) < V_T \sum^k C_{jC}(\sum^k C_{jC} - 2C_{TC}). \tag{8}$$

The right side of (8) will always be negative (all $C_{jC} \geq 0$), and it can be seen that items for which the left side of (8) is positive will not satisfy (3) or the subsequent conditions. For since the terms other than the numerator on the right in (6) are positive, the k items (including the case $k = 1$) may not be discarded by (6) with a resulting increase in validity unless $S_T > S_{T-k}$. But $V_k - 2C_{kT}$ in (8) is equal to $V_{T-k} - V_T$, a negative quantity when $S_T > S_{T-k}$. Multiplying (8) by -1 and rearranging terms, we obtain

$$\frac{C_{TC}^2}{V_T} > \frac{\sum^k C_{jC}(2C_{TC} - \sum^k C_{jC})}{2C_{kT} - V_k}, \qquad 2C_{kT} > V_k. \tag{9}$$

Although (9) is the general condition (under the restriction, $C_{jC} \geq 0$, from which it follows that $2C_{kT} > V_k$) for rejecting k items from T in order to obtain a more valid test, the case where $k = 1$ is more useful. When k is a single item j, (9) reduces to

$$\frac{C_{TC}^2}{V_T} > \frac{C_{jC}(2C_{TC} - C_{jC})}{2C_{jT} - V_j}, \qquad 2C_{jT} > V_j. \tag{10}$$

The second Gleser and DuBois condition referred to previously is a fairly good approximation of (10). It can be obtained from (7) when k is a single item j [another way of writing (10)] by dividing (7) by S_T and substituting $S_T S_{T-j} \doteq (V_T + V_{T-j})/2$; (7), (9), and (10) have recurrence properties not possessed by any approximations known to the writer.

A proof follows that each successive application of (7) when $k = 1$, that is, when single items are discarded successively, increases validity. It can be generalized immediately for $k > 1$. The first inequalities in the recursion series for (7), when single items 1, 2, $\cdots$ are rejected, are

$$\frac{C_{TC}}{S_T} > \frac{C_{1C}}{S_T - S_{T-1}}, \qquad \frac{C_{(T-1)C}}{S_{T-1}} > \frac{C_{2C}}{S_{T-1} - S_{T-2}}, \qquad \cdots. \tag{11}$$

If the terms on the left of the inequality signs in sequence (11) were divided by the constant S_C, they would become the correlations which are the validity coefficients as the test becomes shorter. These terms therefore comprise a

sequence with a maximum upper bound equal to S_C . Every bounded monotone sequence is convergent, and

$$C_{TC}/S_T , \qquad C_{(T-1)C}/S_{T-1} , \qquad \cdots \tag{12}$$

is bounded and can be shown to be monotone. Given that the first inequality of (11) holds for item 1, we first prove that

$$C_{TC}/S_T < C_{(T-1)C}/S_{T-1} . \tag{13}$$

By substituting $C_{(T-1)C} = C_{TC} - C_{1C}$ and clearing of fractions, (13) can also be written

$$S_{T-1}C_{TC} < S_TC_{TC} - S_TC_{1C} . \tag{14}$$

But this is another way of writing the first term of (11), and this proves that discarding item 1 has increased validity from r_{TC} to $r_{(T-1)C}$. By induction, if subsequent items are rejected by (11), the validity always increases. The sequence (12) is therefore monotone and convergent. By comparison with (12) the sequence

$$\frac{C_{1C}}{S_T - S_{T-1}} , \qquad \frac{C_{2C}}{S_{T-1} - S_{T-2}} , \qquad \cdots , \tag{15}$$

which is formed by the terms on the right of the inequalities in sequence (11), is also convergent but not necessarily monotone. Conditions can be written which will make (15) monotone, but the writer was unable to find any which were practical and at the same time would insure unique maximum validity.

Nevertheless, (11) converges as it stands. (A difference between two convergent sequences also converges). This is sufficient reason to make use of an analogous sequence, one comprised of terms which are approximations of (10), as an aid in constructing valid tests.

II. *Development of a Practical Test Construction Procedure*

Since rejection of a unique set of items from a test in order to maximize (obtained) validity for a given sample is in general not possible, and since the sampling variance of the validity when both items and persons are sampled is unknown, the need for item selection restrictions in addition to (10) is obvious.

A practical restriction traditionally used in test construction is that items should each have enough variance to share appreciably in the discrimination of subjects. A second restriction, that items retained in the test have individual validity, follows from (6). Reference to (6) when k is a single item j shows that because items are scored so that $C_{jC} \geq 0$, discarding items for which $2C_{jT} - V_j \equiv V_T - V_{T-j} < 0$ will not increase test validity. Because of the scoring convention, therefore, the sign of $S_T - S_{T-j}$ is de-

pendent upon the sign of C_{ic} , which in turn is subject to sampling fluctuations. In order to be reasonably confident that these signs, and consequently the test scoring, will not vary in subsequent samples, it is necessary that most of the items in the original test be valid. This condition alone could be satisfied by requiring that each item-criterion correlation significantly exceed zero, or that

$$C_{ic} \geq S_i S_c \, z/\sqrt{N}, \tag{16}$$

where z is a chosen critical value for a normal deviate, say 1.96.

Setting $S_i = 0.5$, the maximum value, in (16) will provide a conservative test for all items for which $S_i < 0.5$. That is, items satisfying (16) when $S_i = 0.5$ are significantly correlated with C at the level specified by z, but if a low-variance item is to be included in the test, then this must be compensated for by a higher covariance with the criterion. Including in the initial test T only items satisfying (16) when $S_i = 0.5$, therefore, has the desired effect of insuring both that the initial test will contain valid items, and that most of them will have large variances. Rejection of items from T can then be started by applying an approximation of (10).

Ideally (10), or an approximation, would be recomputed after any item had been discarded by it before discarding another item. But rejecting items one at a time is not feasible because of the labor of recomputing values for the C_{iT} . A workable procedure, which is essentially the same as that used by Gleser and DuBois, is first to reject *all* items for which (10) holds (without recomputing C_{iT}) and then to reexamine the rejected items to see which would logically be put back into the test. The new test so formed may then be treated as if it were an initial test and the process continued until there are no further increases in validity.

In order to secure a suitable approximation of (10), first note that C_{ic}^2 is the final term in the expanded numerator on the right, and that C_{Tc} , V_i , and C_{ic}^2 are in descending orders of magnitude. Substituting for the final C_{ic} its mean value in T, $C_{ic} \doteq C_{Tc}/n$, and substituting the *maximum* value for V_i , 0.25, reduces (10) to

$$\left(\frac{2n}{2n-1}\right)\frac{C_{Tc}}{V_T} > \frac{C_{ic}}{C_{iT} - 0.125}, \tag{17}$$

where n is the number of items in T. The use of the approximation $V_i \doteq 0.25$ is justified because, first, after applying (16) as described, most item variances are large; second, marginal items which would ordinarily be retained by (10) despite small variances will not be discarded by (17); and finally, the exactness of the value for V_i in (17) is increasingly unimportant with increasing n.

Suppose that m items for which (17) holds have been discarded, leaving a residual test M. Since (17) is progressively less accurate as items beyond the first are discarded, it is likely that some of the discarded items should be

put back into the test. Any rejected item j which satisfies the validity condition

$$r_{MC} \leq r_{(M+j)C} \tag{18}$$

will be added to test M. By a development similar to that used in deriving (10), (18) may be written

$$\frac{C_{MC}^2}{V_M} \leq \frac{C_{jC}(2C_{MC} + C_{jC})}{2C_{jM} + V_j}. \tag{19}$$

Using the approximations for C_{jC}^2 and V_j described above, (19) becomes

$$\left(\frac{2m}{2m+1}\right)\frac{C_{MC}}{V_M} \leq \frac{C_{jC}}{C_{jM} + 0.125}, \tag{20}$$

where m is the number of items in M.

Items for which (20) holds are added to M to form a third test. It is then possible to treat this third test as if it were T, the initial test, and again apply conditions (17) and (20); this process may be continued until no further increase in validity occurs.

It is also desirable to simplify computation of the covariances in (17) and (20). This can be done as follows: Test and criterion distributions are first transformed to comprise only five symmetrical categories containing the 8 per cent highest, 18 per cent next highest, 48 per cent middle, 18 per cent low, and 8 per cent lowest scores. A division of distributions into categories containing 9, 20, 42, 20, and 9 per cents of cases is recommended by Flanagan (5) as maximally efficient when scores in the categories are assigned values 2, 1, 0, − 1, and − 2, respectively. The distribution used in the present paper will be slightly less efficient than the one recommended by Flanagan, but will have the advantage of possessing a unit variance. Using only five categories, the covariances in conditions (17) and (20) may be obtained from item counts for the four extreme categories.

To transform (17) and (20) so that the covariances are of the form

$$C' = (2e + f - g - 2h)/N = D/N, \tag{21}$$

where e, f, g, and h are frequencies in the categories in the order given above (scores with zero weights, which are in the center category, are ignored), first note that the variance of every such forced distribution is a constant, $V' = 1.00$. Then if transformed values are indicated by primes, $S_C' = S_T' = 1$, and

$$r_{TC} = C_{TC}/(S_T S_C) = C'_{TC}/(S_T' S_C') = C'_{TC}. \tag{22}$$

Only one transformed distribution corresponding either to r_{jC} or to r_{jT} is needed; thus it is assumed that

$$r_{jT} = C_{jT}/(S_j S_T) = C'_{jT}/(S_j S_T') = C'_{jT}/S_j, \tag{23}$$

and a similar expression can be written for r_{iC} . Transforming (17) gives

$$\left(\frac{2n}{2n-1}\right)C'_{TC} > \frac{C'_{iC}}{C'_{iT} - (0.125/S_T)}, \tag{24}$$

and transforming (20),

$$\left(\frac{2m}{2m+1}\right)C'_{MC} \leq \frac{C'_{iC}}{C'_{iM} + (0.125/S_M)}. \tag{25}$$

A more convenient form of (24) is obtained by substituting $\sum^N T'C'/N = C'_{TC}$ and, from (21), $D/N = C'$ to get

$$\left(\frac{2n}{2n-1}\right)\frac{\sum^N T'C'}{N} > \frac{D_{iC}}{D_{iT} - (0.125N/S_T)}, \tag{26}$$

where n is the number of items in T. Similarly (25) is conveniently written

$$\left(\frac{2m}{2m+1}\right)\frac{\sum^N M'C'}{N} \leq \frac{D_{iC}}{D_{iM} + (0.125N/S_M)}, \tag{27}$$

where m is the number of items in M. Very little loss of accuracy occurs in (27), especially when $m/n > .8$ and $m > 30$, if the value of S_M is taken to be mS_T/n.

The procedure for applying (16), (26), and (27) is outlined in the next section.

III. *The Procedure for Applying the Item Selection Conditions*

1. Separate the answer sheets into five piles according to their criterion scores C, containing the 8 per cent highest, 18 per cent next highest, 48 per cent middle, 18 per cent low, 8 per cent lowest C scores, respectively. Mark papers in the four extreme C-score categories e, f, g, and h, respectively, so they can be identified in Step 6.

2. Record on item analysis sheets the values for e, f, g, and h, which are the frequencies of a response (for example, "true") in the highest, next highest, low, and lowest C-categories. Obtain from the item counts for the four extreme categories the difference

$$D_{iC} = 2e + f - (g + 2h) = NC'_{iC}$$

for each item j [see (21)]. Next choose the direction of response for each item so that $2e + f \geq g + 2h$, that is, so that $D_{iC} \geq 0$, and mark the items accordingly.

3. Apply (16) by including in the initial test T only those items for which $D_{iC} \geq 0.5z\sqrt{N}$, where z is the normal deviate corresponding to a chosen level of significance.

4. Score test T (with items scored in the directions determined in Step 2) and mark the scores on the answer sheets. Tally T and obtain its standard deviation S_T .

5. Separate the answer sheets into five piles as in Step 1, but this time according to their T scores.

6. Write in the frequencies for the cells of the 5×5 contingency table for the transformed T-scores and C-scores. Compute $\sum^N T'C'$, first noting that scores falling in the most extreme categories on either variable receive values $+2$ or -2, while those in next most extreme categories receive values $+1$ or -1. This can be done in a few minutes by counting, and frequencies for the center categories can be ignored.

7. Obtain the D_{iT} from item counts as in Step 2. Check for retention in the test any items for which $D_{iT} \leq 0.125N/S_T$. It is possible there may be no such items.

8. Compute the constant which is the left side of (26); set up the right side of (26) for each item j (the operations may not have to be carried out) and retain only those items for which (26) fails to hold.

9. Score the test M, comprised of items retained after Step 8. Tally M and obtain its standard deviation S_M , or use $S_M = mS_T/n$.

10. Separate the answer sheets into five piles as in Step 1, this time according to the M scores.

11. Obtain D_{iM} from item counts as in Step 3, *but only for those items previously discarded in Step 8.*

12. Obtain $\sum^N M'C'$ as in Step 6 and compute the left side of (27). Set up the right side of (27) for each item for which D_{iM} was obtained in Step 11, and mark items for which (27) holds to be put back into test M. This completes the first cycle of the iteration, and a large proportion of the possible increase in validity will have been obtained.

13. For convenience, again call the test obtained after Step 12 "test T," and repeat the procedure starting with Step 5. The iteration will stop at a point where either (26) or (27), applied alternately, will produce no further increase in test validity. Always apply (27) to all previously rejected items. If S' is a transformed score on the final test, $\sum^N S'C'/N$ [see (26) and (27)] is a conservative estimate of the final validity coefficient.

IV. *The Problem of Test Homogeneity*

A question which naturally arises is how much internal consistency tests will have when constructed by the method. Since the test becomes shorter and at the same time some of the redundant items with higher test correlations are dropped, its homogeneity may decrease. It will usually remain relatively high, however, especially if the initial test is long enough. Aside from length, another reason why homogeneity will be high (in practice between .82 and .90 for final tests of about 100 items) is that initial test

items are limited by (16) to those having significant criterion covariances; this condition alone tends to select items which correlate positively with the total test and thus enhance homogeneity.

Cronbach (2) has demonstrated the relative importance of high item-test correlations and test length in contributing to homogeneity. In symbolism of the present paper,

$$\frac{C_{jT}}{V_T} > \frac{V_j}{2}\left(\frac{1}{\sum_n V_i} - \frac{1}{V_T}\right), \qquad n > 7, \tag{28}$$

is a close approximation to $r_{(T+j)(T+j)} > r_{TT}$ where the latter are K.R. 20 coefficients for test T, including and excluding item j, respectively. Inequality (28) can be used to determine how large the item-test covariance should be before item j will contribute to the homogeneity of T. When (28) is transformed as in Part II, it is found, for tests and samples of only moderate size, that all items for which $D_{jT} > 1$ will contribute to homogeneity.

V. *Applications and Discussion*

The method is not time-consuming; once it is learned, two persons working with about 300 answer sheets, each of which contains as many as 540 true-false item responses, can construct a test in about 14 hours.

In the first two applications, the criterion variable was the score on an attitude test, the 20-item California Ethnocentrism Scale (1), and the tests constructed to correlate with it were selected from 379 true-false items from various standardized personality inventories. Items in the Ethnocentrism scale are hostile or disparaging statements about minority groups; each item receives a score from 1 to 7 to indicate extent of agreement.

In the first application using a sample of 288 college women, (16) was applied using $z = 1.96$ (see Part III, Step 3) with the result that 79 items were selected to comprise an initial test for which the validity, $\sum T'C'/N$, was .62. Of these, 14 were rejected by (26) leaving a 65-item test with a validity of .64. Applying (27) to the rejected items put one of them back, giving a 66-item test with a validity of .65. Subsequent applications of (26) and (27) resulted in rejections and selections of from 1 to 4 different items at a time accompanied by slight decreases in validity. The 66-item test was therefore accepted as the final test for this sample.

In the second application, using a sample of 50 middle-aged women, only 41 items from the 379 were selected by (16) for the initial test, even though z had been chosen because of the small sample size to correspond to the .10 level of significance. The initial validity for the 41 items was .77, a value undoubtedly largely spurious because of chance item-criterion correlations and because of the small sample size. Application of (26) discarded 9 items to increase validity to .80, and an application of (27) put 6 items back

into the test to make a 38-item test with validity of .82. This was not exceeded by subsequent iterations.

In several applications the necessity for using only items which were individually valid was demonstrated. In one case all positively scored items were used in the initial test, that is, condition (16) was not applied. The convergence was slow, and because of the presence of items which would be invalid in subsequent samples, the gain in test validity would not be expected to be permanent (see Part II). In another case only 58 items out of 677 available could be found which were related at the .05 level to grades of college freshmen. About 34 items would therefore be expected to have only a chance relationship to grades. Application of the method retained 41 items and raised validity from .52 to .60. But in a subsequent sample the validity was almost as small for the shortened test (.22) as for the initial test (.16). The shrinkage in both cases was obviously due to the large proportion of invalid items among the initial 58. These results show that it is necessary to insure item validity in the initial test before applying the rest of the method; this may be achieved either by applying (16) with z large when there is only one large sample, or by using several samples if z must be smaller.

If most of the items are valid, the method appears to be worth applying. For example, in a study reported elsewhere 178 items were found each to correlate at the .01 level with a criterion. The test validity was raised from .66 to .78 in the first sample ($N = 441$) by applying Flanagan's method (5), which is an approximation of the present method. A year later the shortened test (124 items) correlated .74 in a new sample ($N = 402$). In this case the difference .74 − .66 is significant, using the traditional z-transformation test, at the .03 level. Despite this apparently permanent gain in validity, the merit of selecting items for validity cannot be finally assessed until the appropriate sampling statistics are derived and applied.

REFERENCES

1. Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J., and Sanford, R. N. The authoritarian personality. New York: Harper, 1950.
2. Cronbach, L. J. Coefficient α and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.
3. Fisher, R. A. The precision of discriminant functions. *Ann. Eugen., Lond.*, 1940, **10**, 422–429.
4. Flanagan, J. C. A short method of selecting the best combination of test items for a particular purpose. *Psychol. Bull.*, 1936, **33**, 603–4.
5. Flanagan, J. C. The effectiveness of short methods for calculating correlation coefficients. *Psychol. Bull.*, 1952, **49**, 342–348.
6. Gleser, G. C., and DuBois, P. H. A successive approximation method of maximizing test validity. *Psychometrika*, 1951, **16**, 129–139.
7. Guilford, J. P. Psychometric methods (2nd Ed.). New York: McGraw-Hill, 1954.
8. Gulliksen, H. Theory of mental tests. New York: Wiley, 1950. Chap. 21.
9. Horst, P. Item selection by means of a maximizing function. *Psychometrika*, 1936, **1**, 229–244.

10. Hotelling, H. Relations between two sets of variates. *Biometrika*. 1936, **28**, 321–377.
11. Loevinger, Jane, Gleser, Goldine C., and DuBois, P. H. Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, **18**, 309–317.
12. Lord, F. M. Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 1955, **20**, 1–22.
13. Richardson, M. W., and Adkins, Dorothy C. A rapid method of selecting test items. *J. educ. Psychol.*, 1938, **29**, 547–552.
14. Toops, H. A. The *L*-method. *Psychometrika*, 1941, **6**, 249–266.
15. Tukey, J. W. Unsolved problems of experimental statistics. *J. Amer. statist. Assoc.*, 1954, **49**, 706–731.
16. Webster, H. Approximating maximum test validity by a non-parametric method. *Psychometrika*, 1953, **18**, 207–211.

Manuscript received 3/3/55

Revised manuscript received 5/28/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

THE VALIDITY OF THE SUCCESSIVE INTERVALS METHOD OF PSYCHOMETRIC SCALING*

WILLIAM W. ROZEBOOM AND LYLE V. JONES
UNIVERSITY OF CHICAGO

The degree to which scale values computed by the method of successive intervals diverge from theoretically "true" values is seen to be due to three types of error: error due to inequalities in variances of the distributions from which the scale values are computed, error due to non-normality of the distributions, and sampling error. The contribution of each type of error to the total error is evaluated; the latter is seen to be surprisingly small under appropriate conditions. Certain aspects of the formal methodology underlying scaling procedures are also briefly considered.

One of the most popular and perhaps the simplest of all methods by which stimuli can be assigned values for some psychological variable is the rating scale technique. Basically, a rating scale is some set of categories that partition sets of events into mutually exclusive classes. For example, a rating scale might be defined by the categories *high*, *medium*, and *low*, and a set of events generated by "the evaluation of the esthetic value of art object i by judge j," where i and j range over specified classes of art objects and judges. That is, each judge j assigns each art object i to a category of the rating scale, such an assignment constituting an event. Corresponding to each event designated by the coordinate pair i, j there is one and only one category: *high*, *medium*, or *low*.

Usually, the localization of each event on the scale is only a means to a representation of various subclasses of the events by a single value of the scale. This can be done by taking the most representative scale value of the distribution of scores in a subclass as the scale value for the subclass as a whole. Thus, for the example already given, we may be less concerned with the rating given to a particular art object by a given judge than we are in a rating representative of the values assigned to that object by the various judges. Since a specific art object defines a subclass of ratings, the most representative rating (however defined) can be taken as the value of this stimulus on the esthetic scale, not dependent upon any particular judge.

In its most elementary form, a rating scale imparts no measure of quantity to the events rated by it, merely being comprised of a set of mutually exclusive

*This paper reports research undertaken in cooperation with the Quartermaster Food and Container Institute for the Armed Forces, and has been assigned number 475 in the series of papers approved for publication. The views or conclusions contained in this report are those of the authors. They are not to be construed as necessarily reflecting the views or indorsement of the Department of Defense.

categories. An example of this simplest type is the standard color chart by which colors are classified. Stevens (9) has named this kind of scale a *nominal* scale. With appropriate additional data, assumptions, or definitions, however, the rating scale can be utilized as an ordinal, or even an interval scale.

If a relation can be obtained which orders the categories, then the rating scale has become an ordinal scale for that relation. One of the more customary ordering relations employed by psychologists in generating ordinal scales is that of preference, or choice. If the categories are such that, for a given judge, (*i*) an item assigned to category A is always chosen over any item assigned to category B (at least at the time of the assignment) and any item assigned to category B is always chosen over an item assigned to category C, and (*ii*) no item assigned to C is ever chosen over items assigned to A, then categories A, B, and C are ordered by the relation of preference.

It is customary at this point either to define or to hypothesize the existence of a psychological continuum underlying the categories of the rating scale, such that each category covers a range of the continuum, these ranges being exhaustive, mutually exclusive, and in the same ordinal relation as the corresponding categories. In short, the rating scale is interpreted as a gross technique by which the values of events are estimated on a similar, but much more discriminating underlying scale. Thus, art objects evaluated in terms of a three-category scale are assumed to be much more finely dis-

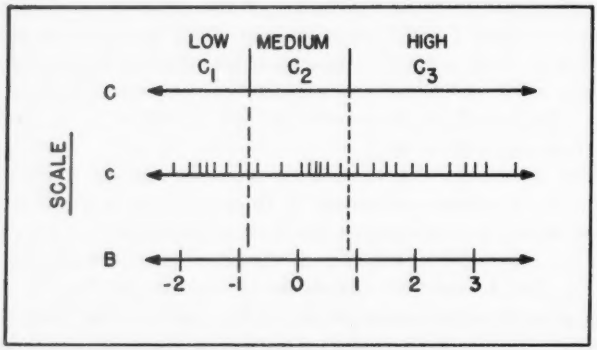


FIGURE 1

Conversion of a Sequence of Ordered Categories into an Interval Scale. Scale C Represents the Observed Rating Categories; Scale c, the Assumed Underlying Continuum; and Scale B, the Metric (with Arbitrary Origin and Unit Distance) Assigned to Scale c.

tinguishable esthetically. This is illustrated in Figure 1. Scale C is comprised of the categories C_1 , C_2 , C_3 . Events falling in a category C_i are ordered as higher or lower than events in $C_j (i \neq j)$ for the property being rated, but no distinction is made among the events falling in C_i . Scale c is the continuum which is hypothesized or defined to underlie scale C, the smaller categories

indicating much finer differences in degree of the rated property. Strictly speaking, the underlying scale need not be an actual continuum; it may be conceptualized as a finite number of subdivisions of each category of the rating scale, as long as these subcategories are ordered by the same relation that orders the rating scale.

If numbers can now be assigned to the various positions on the underlying scale in such a manner that an interpretation can be defined, discovered, or assumed for the relative differences between positions on the scale, the theoretical underlying scale becomes an interval scale indicated in Figure 1 by scale B. Then the assignment of an event to a category of C is interpreted as an estimation of the score of the event on the underlying scale B. We shall refer to theoretical metrics such as B, defined or inferred from cruder empirical measures, as *base scales*. It has been cutomary to define the base scale (more rigorously, a set of base scales, linear transformations of one another) for a particular set of categories and distributions over the categories as that assignment of numbers to the continuum which normalizes the distributions (**10**). There may be other equally valid ways of defining the base scale, e.g., the counting of just-noticeable-differences, or defining the base scale so as to normalize distributions other than the ones being dealt with in the given study. It is not always possible, given more than one distribution over the same continuum, to find a numbering of the continuum that simultaneously normalizes all the distributions. Although the method of successive intervals, as described in the literature, has assumed normality for all distributions used in the analysis, we shall demonstrate that the validity of the method as a computational technique need not assume normal distributions.

Once the existence of a base scale has been defined over the categories of a rating scale, each event classified by the scale is considered to have a value on the base scale. Since each category corresponds to an interval of the base scale, the assignment of an event to a specific category determines a range in which its base scale value falls. If, now, the shape of a distribution of scores on the base scale is known, values for the widths of the various category intervals can be computed in terms of the standard deviation of that distribution as the unit of measurement. These are computed by tabulating the cumulative proportion of scores at each boundary of the interval and calculating the width in sigmas corresponding to such a percentile difference for that type of distribution. If the distribution is known or assumed to be normal, then the interval width will be the standard deviation of the distribution multiplied by the difference between the normal deviates corresponding to the cumulative proportions at the lower and upper boundaries of the interval. If a number of distributions are available, i.e., a group of judges rates a set of stimuli so that each stimulus determines a class of ratings, a number of measures of each interval width will be obtained in terms of the sigmas of the various distributions. If these are pooled,

estimates of the interval widths in terms of a common unit of measurement are obtained. Finally, if the median of a distribution be taken as the base scale score representing the distribution, the exact base scale value of this point can be estimated as follows: observe the cumulative proportion at the boundaries of the interval in which the median falls, compute from the assumed distribution function the proportion of the distance from the lower boundary, multiply this proportion by the interval width, and add the product to the base scale value for the lower boundary.

This and similar techniques for conversion of distributions of scores of a set of rating scale categories into points along an interval scale have been variously described in the literature, most frequently under the title, *the method of successive intervals* (**1**, **2**, **4**, **5**, **6**, **7**, **8**). The general computational steps usually given for evaluation of the interval widths are: (*i*) for each distribution, compute each interval width in terms of the sigma of that distribution by taking the difference between the normal deviates corresponding to the boundaries of the interval, assuming each distribution to be normal; (*ii*) let the average value of the computed widths for a given interval be taken as the best estimate of the width of that interval in terms of a unit of measurement common to all intervals. When the cumulative proportion at a boundary of an interval is nearly 0 or 1, the estimate of interval width given by that distribution for the interval is too unreliable for use, so the average width for an interval must be a weighted average; the weights of 0 or 1 have been employed in all past applications of the method.

Previous advocates of the method of successive intervals have attempted to validate the technique by demonstrating its extremely high correlation with the method of paired comparisions (**8**), and its internal consistency (**3**). It is our present aim to evaluate the method in terms of the degree to which results of the computations from empirical data can be expected to diverge from theoretically "true" values as determined from the definition of the base scale. That is, we propose to evaluate the *absolute* validity of the method.

The primary scores which are determined by the method of successive intervals are the widths of the category intervals relative to some arbitrary unit of measurement. The location of medians of the various distributions is secondary to estimation of the interval widths, since once the latter are known the former are easily determined. It is obvious that (*i*) if all the distributions used to measure the interval widths have equal variances, (*ii*) if all the distributions are normal, and (*iii*) if there are no sampling errors, then the computed values of relative interval widths are identical with the theoretical values. For, if each experimentally obtained distribution were normal, and for every distribution the proportion of cases falling within each interval showed no sampling errors, then the interval widths computed from a given distribution would be identical with the theoretical values as measured by the standard deviation of that distribution. If the variances

of all the distributions were equal, then each distribution would give the same computed value for a given interval width. Thus there are three possible sources of error in computation of base scale values by means of the method of successive intervals: type (*a*) errors due to unequal variances of the distributions used to compute the interval widths, type (*b*) errors due to non-normality of the distributions, and type (*c*) sampling errors, i.e., errors due to the estimation of cumulative proportions of the interval boundaries from finite samples of the measuring distributions.

As a tool for evaluation of the contributions of these sources of error to a total error of estimate of relative interval width, it is convenient to define a *coefficient of error*. Let a quantity χ be estimated by a quantity X. Then the coefficient of error, ξ, for the estimation of χ by X is $\xi = (X - \chi)/\chi$ or $X = (1 + \xi)\chi$. The magnitude of the coefficient of error gives the discrepancy between X and χ as a proportion of χ and is nothing more than 1/100 of the percentage error in the approximation of χ by X.

Relative interval widths computed by the method of successive intervals are estimates of true *relative* interval widths on the base scale. By relative interval widths, we recognize that the unit of measurement is arbitrary, so that the ratio of one interval width to another (which is invariant under transformation of the unit of measurement) is the critical quantity by which relative interval width is expressed. We can evaluate the coefficient of error for the estimation of true interval width ratios from computed ratios as follows: (*i*) find an expression for the computed interval widths, L_j and L_k, for categories j and k, in terms of the three types of errors that influence the computed widths, and of the true interval widths, λ_j and λ_k; (*ii*) set

$$L_j/L_k = (1 + \xi_{jk})(\lambda_j/\lambda_k). \tag{1}$$

Solving for ξ_{jk}, we obtain the coefficient of error for the estimation of relative interval widths by the method of successive intervals as a function of the different types of error; we shall be able to see explicitly the manner and extent to which each kind of error contributes to the total error.

Let λ_j be the true width on the base scale of an interval j in terms of some arbitrary unit of measurement U, and L_j the width of the interval as computed by the method of successive intervals. Let η_i (measured in terms of U) be the standard deviation of the ith distribution over the base scale. If the ith distribution is normal and displays no sampling errors, then the cumulative proportions at the upper and lower boundaries of interval j permit an exact computation (through use of a table of the normal proability integral) of the magnitude of λ_j in terms of η_i as a unit of measurement. Specifically,

$$l_{ij} = \lambda_j/\eta_i = \nu_i\lambda_j ,$$

where l_{ij} is the width of interval j as computed from distribution i; λ_j and η_i are the true magnitudes of interval j and the standard deviation of distri-

bution i, respectively, in terms of the arbitrary unit of measurement U; and $\nu_i = 1/\eta_i$. However, to the extent that distribution i diverges from normality and contains sampling errors, l_{ij} will differ from $\nu_i\lambda_j$. In general, $l_{ij} = \nu_i\lambda_j + \zeta_{ij}$, where ζ_{ij} is the discrepancy between the computed l_{ij} and the theoretical $\nu_i\lambda_j$. ζ_{ij} can be analyzed into two additive components E_{ij} and ϵ_{ij}, where E_{ij} is a *constant* bias due to non-normality of the distribution and ϵ_{ij} is a random sampling error. Thus,

$$l_{ij} = \nu_i\lambda_j + E_{ij} + \epsilon_{ij} . \tag{2}$$

It should be noted that the unit of measurement for the error terms E_{ij} and ϵ_{ij} is η_i, the standard deviation of the ith distribution; while the unit of measurement for $1/\nu_i$ and λ_j is the arbitrary U, which is the same for all distributions.

The computed width L_j of interval j is a weighted average of the estimates of widths contributed by the various distributions. That is,

$$L_j = \sum_i w_{ij}l_{ij} = \lambda_j \sum_i w_{ij}\nu_i + \sum_i w_{ij}E_{ij} + \sum_i w_{ij}\epsilon_{ij} , \tag{3}$$

where $\sum_i w_{ij} = 1$. Defining the quantities A_j, β_j, and γ_j by

$$A_j = \sum_i w_{ij}\nu_i , \tag{4}$$

$$\beta_j = (\sum_i w_{ij}E_{ij})/(A_j\lambda_j), \tag{5}$$

$$\gamma_j = (\sum_i w_{ij}\epsilon_{ij})/(A_j\lambda_j), \tag{6}$$

we obtain

$$L_j = A_j\lambda_j(1 + \beta_j + \gamma_j). \tag{7}$$

Since λ_j is inversely proportional to, and A_j proportional to the magnitude of the base unit of measurement U, $A_j\lambda_j$ is invariant for transformations of U. Therefore β_j and γ_j, which are also invariant under transformations of U, may be interpreted as error per unit length of interval due to non-normality of distribution and to sampling error, respectively. L_j may be interpreted as an estimate of λ_j with $1/A_j$ as the unit of measurement. It will be seen below that $1/A_j$ is approximately the harmonic mean of the standard deviations of the measuring distributions.

We are now able to evaluate the coefficient of error, ξ_{jk}, for the computed ratio L_j/L_k as an estimate of the true ratio λ_j/λ_k of the widths of intervals j and k. Finding L_k by substitution of k for j throughout (7) and solving for ξ_{jk} in (1) we find that

$$\xi_{jk} = \frac{A_j}{A_k}\left(\frac{1 + \beta_j + \gamma_j}{1 + \beta_k + \gamma_k}\right) - 1,$$

which may be written

$$\xi_{jk} = \alpha_{jk}\left(\frac{1 + \beta_j + \gamma_j}{1 + \beta_k + \gamma_k}\right) + \frac{\beta_j - \beta_k}{1 + \beta_k + \gamma_k} + \frac{\gamma_j - \gamma_k}{1 + \beta_k + \gamma_k}, \tag{8}$$

where

$$\alpha_{jk} = (A_j/A_k) - 1. \tag{9}$$

Since α_{jk} reflects the difference between units of measurement within interval j and within interval k, and vanishes (as shown below) when the variances of all the distributions are equal, α_{jk} may be regarded as the error in relative interval width due to unequal variances of the distributions which are used in estimating the interval widths. Thus, of the three sources of error in the method of successive intervals, type (a) is represented quantitatively by α, type (b) by β, and type (c) by γ.

γ-Error

It will be recalled from (3) that each distribution i was assigned a weight w_{ij} for its contribution l_{ij} in the computation of L_j. It is now possible to assign these weights in a manner that minimizes the sampling error γ_j. Assuming the various distributions to be essentially independent of one another in their sampling errors, we find from (6) the mean and variance for γ_j (under repeated sampling with a fixed set of weights) to be

$$\mu_{\gamma j} = \sum^{i} w_{ij}\mu_{\epsilon ij}/A_j\lambda_j ; \tag{10}$$

$$\sigma^2_{\gamma j} = \sum^{i} w^2_{ij}\sigma^2_{\epsilon ij}/(A_j\lambda_j)^2. \tag{11}$$

But $\epsilon_{ij} = \delta_{Uij} - \delta_{Lij}$, where δ_{Uij} and δ_{Lij} are the sampling errors for the standardized deviates of the normal probability distribution corresponding to the cumulative proportions of distribution i at the upper and the lower boundaries of interval j, and $\delta \simeq \Delta P/y$, where ΔP is the sampling error of a cumulative proportion at an interval boundary and y is the ordinate of the normal probability distribution at P. Since the sampling mean of ΔP is zero, $\mu_{\epsilon ij} \simeq 0$ and hence, from (10),

$$\mu_{\gamma j} \simeq 0; \tag{12}$$

while

$$\begin{aligned}\sigma^2_{\epsilon ij} &= \sigma^2_{\delta Uij} + \sigma^2_{\delta Lij} - 2\,\mathrm{cov}\,(\delta_{Uij}, \delta_{Lij}) \\ &\simeq \frac{1}{n_i}\left[\frac{P_{Uij}(1 - P_{Uij})}{y^2_{Uij}} + \frac{P_{Lij}(1 - P_{Lij})}{y^2_{Lij}} - \frac{2P_{Lij}(1 - P_{Uij})}{y_{Lij}y_{Uij}}\right],\end{aligned} \tag{13}$$

where n_i is the sample size of distribution i, P_{Lij} (P_{Uij}) is the parametric cumulative proportion of distribution i falling at the lower (upper) boundary

of interval j, and $y_{L_{ij}}$ ($y_{U_{ij}}$) is the ordinate of the normal probability distribution at $P_{L_{ij}}$ ($P_{U_{ij}}$). Since the sample size n_i is known, and the sample cumulative proportions provide sufficiently close approximations to the parametric cumulative proportions, very close approximations of $\sigma^2_{\epsilon_{ij}}$ may be computed from empirical data by use of formula (13). Furthermore, $\sigma^2_{\epsilon_{ij}}$ may be made as small as desired by choosing the sample size n_i sufficiently large.

The assumption of (11) might at first seem gratuitous; for many experimental situations the sampling errors of one distribution will not be strictly independent of the next. Thus, if two sample distributions are obtained from judgments for two stimuli by the same judges, the sampling errors of the two distributions would probably be correlated. However, the disturbing effects of such a lack of strict independence are vitiated by the following considerations: (*i*) factors linking the sampling errors of two distributions usually comprise only a small portion of the total factors determining the outcome of the observed cumulative proportions; (*ii*) linear correlations among the sampling errors may be negligible even when significant non-linear correlations exist; and (*iii*) the intercorrelations may assume both positive and negative values, so that even when their absolute magnitudes are significant their net effect may be negligible. Thus the assumption of (11) involves little loss of generality.

Since by (12) the average value of γ_j is approximately zero, the expected absolute magnitude of γ_j is less than (though on the order of) σ_{γ_j}, so the expected (absolute) size of γ_j will be minimal when σ_{γ_j} is minimal. By differentiation of (11) it will be found that σ_{γ_j} is minimal when, for each i, $w_{ij}\sigma^2_{\epsilon_{ij}} = k_j$, where k_j is a constant of proportionality. Since $\sum_i w_{ij} = 1$, $k_j = (\sum_i \sigma^{-2}_{\epsilon_{ij}})^{-1}$, so

$$w_{ij} = (\sigma^2_{\epsilon_{ij}} \sum_i \sigma^{-2}_{\epsilon_{ij}})^{-1}. \tag{14}$$

Equations (14) and (13) provide the steps for computation of the proper weights. (Except for those distributions for which the cumulative proportion at one of the boundaries of an interval is close to 0 or 1, the weights assigned to the various distributions for that interval are very similar. Hence, the customary procedure of giving zero weight to those distributions for which the sampling reliability of the interval estimate is small and of giving the remaining distributions equal weight in the computation of the interval width should be acceptable for most purposes.)

Substitution of (14) in (11) gives

$$\sigma^2_{\gamma_j} = (A_j\lambda_j)^{-2}(\sum_i \sigma^{-2}_{\epsilon_{ij}})^{-1}. \tag{15}$$

Since $\sigma^2_{\epsilon_{ij}}$ is usually on the order of $1/n_i$, letting n be the average size of the sample distributions and N the number of distributions, $\sigma^2_{\gamma_j}$ is roughly on

the order of $(A_i\lambda_i)^{-2}$ $(nN)^{-1}$. Thus the expected order of magnitude for γ_i is roughly $(A_i\lambda_i)^{-1}$ $(nN)^{-\frac{1}{2}}$. This value may be made as small as desired by taking sufficiently large n and N. For example, if $A_i\lambda_i = .5$, $N = 50$ and $n = 500$, then the expected order of magnitude for γ_i is 10^{-2}. Thus for an empirical study of any substantial proportions an expected order of magnitude for γ_i of 10^{-2} should not be difficult to obtain. In order to maintain a fixed order of magnitude for γ_i , a decrease of interval widths must be compensated for by an increase in (*i*) the sample sizes, (*ii*) the number of distributions, or (*iii*) both. For with σ_{γ_i} held constant, $\sqrt{nN}$ is inversely proportional to $A_i\lambda_i$, while the latter, as shown below, is the width of category j in units of measurement given by the harmonic mean of the standard deviations of the measuring distributions. This has direct implications for the design of rating scales, for it shows that the number of categories into which a scale can be reliably decomposed is limited by the number of stimuli and the size of the population upon which the scale is to be standardized.

Thus, if the width of an interval relative to $1/A_i$ is not too small, and if the study by which the scale is being standardized is of reasonably substantial dimensions, the error in estimation of λ_i due to sampling will be insignificant—generally on the order of 10^{-2}. In light of this, the mean and sampling error of ξ_{ik} can be evaluated. Any reciprocal, $1/s_i$, from a distribution of s with mean M_s can be replaced by the expression $(2/M_s - s_i/M_s^2)$ with an error coefficient of $-[(s_i - M_s)/M_s]^2$. Since from (12), the mean of $(1 + \beta + \gamma)$ is $(1 + \beta)$, $[1/(1 + \beta_k + \gamma_k)]$ may be replaced by $[(1 + \beta_k - \gamma_k)/(1 + \beta_k)^2]$, with an error coefficient of $-[(\gamma_k)/(1 + \beta_k)]^2$, the error of the replacement being negligible so long as β_k does not approach -1. With this replacement we find from (8) and (12) that the sampling mean of ξ_{ik} is

$$\mu_{\xi_{ik}} \simeq \alpha_{ik}[(1 + \beta_i)/(1 + \beta_k)] + (\beta_i - \beta_k)/(1 + \beta_k), \tag{16}$$

while, disregarding second-order terms,

$$\sigma_{\xi_{ij}} \simeq [(1 + \alpha_{ik})/(1 + \beta_k)^2]\sqrt{(1 + \beta_k)^2\sigma_{\gamma_i}^2 + (1 + \beta_i)^2\sigma_{\gamma_k}^2}\,. \tag{17}$$

α-Error

For evaluation of α_{ik} , the error due to inequality of variances, it is convenient to employ the identity

$$A_j = \sum_{i=1}^{N} w_{ij}v_i = N\sigma_{w_j}\sigma_v r_{w_j v} + N\bar{w}_j\bar{v}$$

$$= N\sigma_{w_j}\sigma_v r_{w_j v} + \bar{v},$$

where N is the total number of distributions, σ_{w_j} is the standard deviation of the weights for the jth interval over the N distributions, σ_v is the standard deviation of the v_i over the N distributions, $r_{w_j v}$ is the product-moment

correlation between w_{ij} and v_i over the N distributions, and $\bar{v}$ is the mean value of v_i over the N distributions. From (9), this gives for α_{jk}

$$\begin{aligned}\alpha_{jk} &= \frac{N\sigma_{w_j}\sigma_v r_{w_j v} - N\sigma_{w_k}\sigma_v r_{w_k v}}{N\sigma_{w_j}\sigma_v r_{w_k v} + \bar{v}} \\ &= \left(\frac{C_j r_{w_j v} - C_k r_{w_k v}}{C_k r_{w_k v} V_v + 1}\right) V_v ,\end{aligned} \tag{18}$$

where $C_j = N\sigma_{w_j}$, $C_k = N\sigma_{w_k}$, and $V_v = \sigma_v/\bar{v}$.

The value of C_j depends only on the shape of the distribution of weights for the interval j; this value will be of an order higher than 10^{-1} only when a relatively small proportion of the distributions receive a significant weight for interval j. In the case where a proportion, k, of the distributions receive equal weights and the rest receive 0 weight, $C = \sqrt{(1/k) - 1}$, which exceeds 1 only when $k < .5$ and is no larger than 3 when $k = .1$. The correlation, $r_{w_j v}$, between the weights assigned to the distributions for interval j and the reciprocals of the standard deviations of the distributions can be expected to assume some small negative value (with a chance divergence which vanishes as N grows large), since as η_i increases the boundaries of the interval draw closer to the center of the distribution, yielding an increase in w_{ij} . However, we should expect this correlation to be equal for both intervals j and k. Thus, the maximum value of α_{jk} would be approximately $V_v \times 10^{-1}$.

But V_v is the coefficient of variation for the reciprocals of the standard deviations of the measuring distributions and is approximately equal to the coefficient of variation for the η_i . We shall, as a rule, expect to find V_v on the order of 10^{-1}, which makes α_{jk} on the order of 10^{-2}. Thus only when the variances of the distributions by which the interval widths are computed differ widely among themselves is the error contributed by the inequality of variances of any significance. In such cases, the data can be reanalyzed using the correction for inequalities in variance suggested by Attneave (1).

It should be noted that if all distributions receive equal weights for two intervals j and k, then $\alpha_{jk} = 0$, regardless of the magnitude of V_v . Even when the η_i differ widely, α_{jk} will be negligible if the w_{ij} and w_{ik} are sufficiently homogeneous. It should also be noted that

$$A_j = (C_j r_{w_j v} V_v + 1)\bar{v} \simeq \bar{v}, \tag{19}$$

where $\bar{v}$ is the reciprocal of the harmonic mean of the η_i . This substantiates our earlier contention that the computed interval widths are expressed in units of measurement determined by the harmonic mean of the standard deviations of the measuring distributions.

β-Error

Of the three sources of error in the method of successive intervals, evaluation of β-error is the most difficult. We can replace $\sum_{i=1}^{N} w_{ij}E_{ij}$ by $(N\sigma_{w_j}$

$\sigma_{E_i} r_{w_i E_i} + \bar{E}_i$), where $\bar{E}_i$ and σ_{E_i} are the mean and standard deviation of the errors introduced into the estimation of λ_i by the non-normality of the N distributions. Then from (5)

$$\beta_i = (C_i \sigma_{E_i} r_{w_i E_i} + \bar{E}_i)/(A_i \lambda_i). \tag{20}$$

Note that E_{ij} is measured in terms of the standard deviation, η_i , of distribution i. In particular, when $P_{L_{ij}}$ and $P_{U_{ij}}$ are the cumulative proportions of distribution i at the lower and upper boundaries of interval j, E_{ij} is the difference between the number of sigmas spanned between $P_{L_{ij}}$ and $P_{U_{ij}}$ by the actual distribution i and the number of sigmas spanned between $P_{L_{ij}}$ and $P_{U_{ij}}$ by a normal distribution. Let D_{ij} be the number of sigmas spanned by distribution i between these two cumulative proportions, and let d_{ij} be the corresponding number of sigmas spanned by a normal distribution. Then $E_{ij} = d_{ij} - D_{ij} = \omega_{ij} D_{ij}$, where $\omega_{ij} = (d_{ij}/D_{ij}) - 1$ and is thus the coefficient of error for the approximation of the distance in sigmas spanned between $P_{U_{ij}}$ and $P_{L_{ij}}$ by distribution i by the corresponding distance spanned by a normal distribution. Since $D_{ij} = \lambda_j/\eta_i = \nu_i \lambda_j$, (20) may be rewritten as

$$\begin{aligned} \beta_i &= (C_i r_{w_i E_i} \sigma_{\omega_i \nu \lambda_i} + \overline{\omega_i \nu \lambda_i})/(A_i \lambda_i) \\ &= C_i r_{w_i E_i} \sigma_{\omega_i (\nu/A_i)} + \overline{\omega_i (\nu/A_i)}. \end{aligned}$$

But when V_ν is small,

$$\sigma_{\omega_i (\nu/A_i)} \simeq \sigma_{\omega_i}$$

and

$$\overline{\omega_i (\nu/A_i)} \simeq \sigma_{\omega_i} V_\nu r_{\omega_i \nu} + \bar{\omega}_i \ .$$

Therefore,

$$\begin{aligned} \beta_i &\simeq \sigma_{\omega_i} (C_i r_{w_i E_i} + V_\nu r_{\omega_i \nu}) + \bar{\omega}_i \\ &\simeq \beta_i' + \bar{\omega}_i \ , \end{aligned} \tag{21}$$

where

$$\beta_i' = \sigma_{\omega_i} (C_i r_{w_i E_i} + V_\nu r_{\omega_i \nu}). \tag{22}$$

In general, while there may be some small non-linear correlation between w_{ij} and E_{ij} , the linear $r_{w_i E_i}$ will be close to zero as N increases and the chance fluctuation of $V_{w_i E_i}$ thus diminishes. A similar argument holds for $r_{\omega_i \nu}$; because of the small expected values of C_i and V_ν , β_i' should be on the order of $\sigma_{\omega_i} \times 10^{-1}$ at maximum. It will be shown below that even when a distribution is markedly non-normal, the expected order of magnitude for ω_{ij} is only 10^{-1}, so σ_{ω_i} will be on the order of 10^{-1} at maximum. Thus, β_i' will be on the order of 10^{-2} at maximum and is more likely to be of order 10^{-3}.

It follows that the only likely significant component of β_j is $\bar{\omega}_j$, the latter comprising the average value of ω_{ij} over the N distributions used to measure the width of interval j. These N distributions may be conceived as a sample of size N from an infinite population of potential distributions over the scale. Then $\bar{\omega}_j$ has a sampling mean, $\mu_{\bar{\omega}j}$, and variance, $\sigma^2_{\bar{\omega}j}$, of its own. Similarly, the ω_{ij} for the infinite potential population of distributions over interval j have a mean, μ_{ω_j}, and variance, $\sigma^2_{\omega_j}$. Finally, since $\bar{\omega}_j$ is the mean of a sample of size N from the ω_{ij}, $\mu_{\bar{\omega}j} = \mu_{\omega_j}$ and $\sigma^2_{\bar{\omega}j} \simeq \sigma^2_{\omega_j}/N_j$. (More generally, $\sigma^2_{\omega_j}/N \leq \sigma^2_{\bar{\omega}j} \leq \sigma^2_{\omega_j}$, depending upon the extent to which the ω_{ij} for the sample of N distributions are independent of one another. In most situations, we will expect to find that the ω_{ij} are not wholly independent, but, for the same reasons advanced to justify equation (11), we shall expect that the sum of the covariances will be negligible.) Let β_j'' be the extent to which $\bar{\omega}_j$ diverges from its mean. Then

$$\bar{\omega}_j = \beta_j'' + \mu_{\omega_j}, \tag{23}$$

and thus, from (21),

$$\beta_j = \beta_j' + \beta_j'' + \mu_{\omega_j}. \tag{24}$$

Since β_j'' is of order $\sigma_{\omega_j}/\sqrt{N}$, and as already mentioned, the expected order of magnitude for σ_{ω_j} is 10^{-1} or smaller, then if N is reasonably large the maximum expected order of magnitude for β_j'' is 10^{-2}. This leaves μ_{ω_j} in (21) as the only component of β_j likely to be significant. But μ_{ω_j} is merely the expected value of ω_{ij} on the interval j. As indicated below, the absolute magnitude of ω_{ij} is only of expected order 10^{-1} even when the distributions are quite non-normal. While it is impossible to make any definite statement about the average, μ_{ω_j}, for an interval j, it would seem unlikely that it could exceed .10 except under cases of extreme, persistent, and positively correlated non-normalities among the population of distributions over interval j. Thus, except under unusual circumstances, β_j is of expected order of magnitude 10^{-1} or less, and we may simplify (16) and (17) to

$$\begin{aligned} \mu_{\xi_{jk}} &\simeq \alpha_{jk} + \beta_j - \beta_k \\ &\simeq \alpha_{jk} + (\beta_j' - \beta_k') + (\beta_j'' - \beta_k'') + \mu_{\omega_j} - \mu_{\omega_k} \end{aligned} \tag{25}$$

and

$$\sigma_{\xi_{jk}} \simeq \sqrt{\sigma^2_{\gamma_j} + \sigma^2_{\gamma_k}}. \tag{26}$$

Of the terms in (25), only μ_{ω_j} and μ_{ω_k} are of expected order larger than 10^{-2}.

It yet remains to determine the anticipated order of magnitude for ω. Since the population of potential distributions over a rating scale cannot be specified, it is impossible to assign a mathematical expectation to this term. However, ω may be computed as a function of the degree of non-normality of the distribution being approximated. One may then select a range of distributions within which an empirically encountered distribution reasonably may be anticipated to fall and hence obtain reasonable bounds

for the magnitude of ω. What we shall illustrate here is a technique by which a distribution of any given shape readily may be inspected for its values of ω. By this technique, the reader may select what he considers to be fair examples of empirically anticipated non-normal distributions and easily convince himself that ω is unlikely to be of an order greater than 10^{-1}.

It will be recalled that $\omega = (d/D) - 1$, where d and D are the distances, measured in terms of the standard deviations of the distributions, spanned between the cumulative frequencies at the upper and lower boundaries of the interval by a normal distribution and the empirical distribution, respectively. Let P_U and P_L be the cumulative proportions at the upper and lower boundaries of the interval, let $y(x)$ and $Y(x)$ be the ordinates at x of the unit normal distribution and of the empirical distribution standardized to $\sigma = 1$, respectively, and let x_P and X_P be the distance of cumulative proportion P from the means of the unit normal distribution and the standardized empirical distribution, respectively. Then $d = x_{P_U} - x_{P_L}$ and $D = X_{P_U} - X_{P_L}$. But

$$P_U - P_L = \int_{x_{P_L}}^{x_{P_U}} y(x)\, dx = \bar{y}d,$$

where $\bar{y}$ is the mean value of the ordinate to the unit normal distribution over the interval. Similarly,

$$P_U - P_L = \int_{X_{P_L}}^{X_{P_U}} Y(x)\, dx = \overline{Y}D.$$

Hence $d/D = \bar{Y}/\bar{y}$, so $\omega = (\bar{Y}/\bar{y}) - 1$ and is thus the coefficient of error for the approximation of the average height of the unit normal distribution between two cumulative proportions by the corresponding average height of the standardized empirical distribution. The magnitude of ω is then readily seen by an inspection of the graphs of y and Y against P. That is, let $y(P) = y(x_P)$ and $Y(P) = Y(X_P)$. It is computed without difficulty that $\bar{y}(x) = \tilde{y}(P)$, where $\tilde{y}(P)$ is the harmonic mean of $y(P)$ between P_L and P_U, and similarly $\bar{Y}(X) = \tilde{Y}(P)$. Also, except for those intervals over which the coefficient of variation for $y(P)$ or $Y(P)$ is large, $\tilde{y}(P) \simeq \bar{y}(P)$ and $\tilde{Y}(P) \simeq \bar{Y}(P)$. Thus, given any empirical distribution, the magnitude of the approximation error can be determined readily by standardizing the distribution to unit variance, graphing the height of the distribution against cumulative proportion, and superimposing the corresponding graph of the unit normal distribution. One may then select two cumulative proportions, estimate the average difference between the curves over the interval visually, and divide this by the estimated average ordinate of the normal distribution over the interval.

We illustrate the method through its application to two arbitrary distributions, a rectangular distribution and a triangular distribution skewed

so that the projection of the apex divides the base in a ratio of 1:3. These are shown with unit variance in Figure 2, together with the unit normal distribution by which they are to be approximated. Both distributions represent departures from the normal that, in an empirical distribution, would be considered severe. Figure 3 shows the same distributions in terms

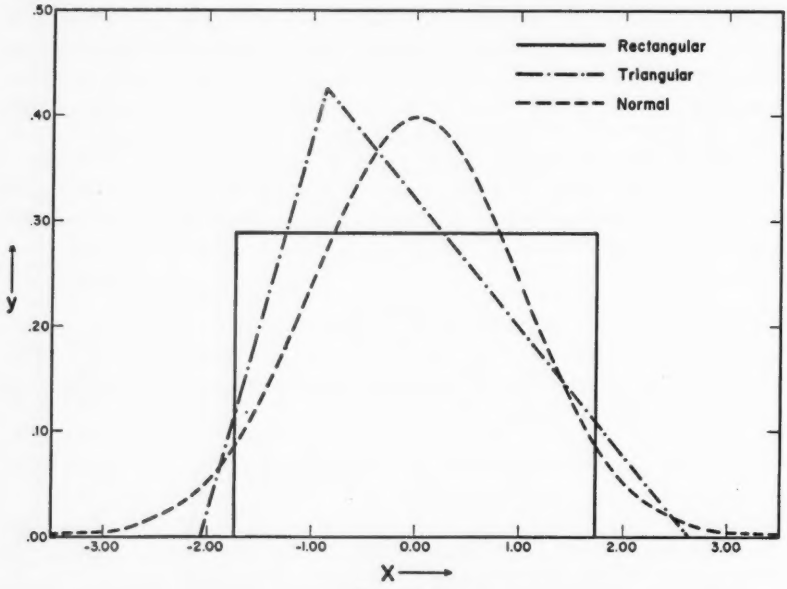


FIGURE 2
Normal, Rectangular, and Triangular Probability Distributions for Which $\mu = 0$, $\sigma = 1$. Projection of Triangular Apex Divides the Base into 1:3 Ratio.

of the ordinates of Figure 2 plotted against the corresponding cumulative proportions. If various pairs of cumulative proportions are selected and ω estimated, it is seen that $|\omega|$ has a modal value in the range.25 to .30 for the two distributions and grows much larger than this only when one of the proportions approaches 0 or 1 (due, here, to the finite ranges of both illustrative distributions). This is typical for most distributions; ω is likely to exceed the order of 10^{-1} only when one of the cumulative proportions at an interval boundary approaches the upper or lower limit. But it is precisely in this case that the error variance of a proportion obtained through finite sampling becomes so large as to give negligible weight to the contribution to the total estimate by an interval width estimate based on such a proportion.

In those few cases where an empirical distribution is likely to show large approximation errors (such as the case of multimodal distribution in which

the modes are well separated and the intervening troughs deep) the severe non-normality of the distribution should be painfully apparent when the distribution is plotted on the successive intervals scale as finally computed. The non-normal distribution then may be discarded and a new analysis of the remaining data performed if the investigator sees fit.

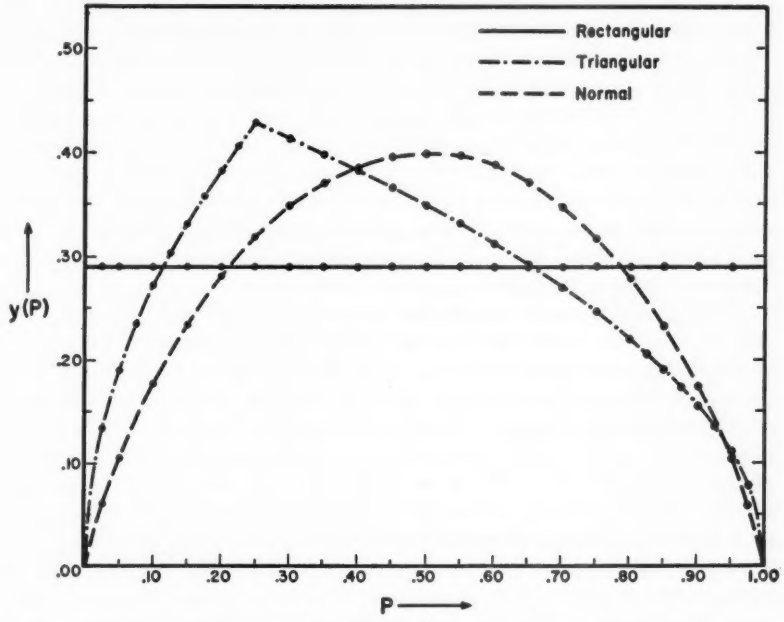


FIGURE 3
Ordinates of the Distributions of Figure 2 as Function of Cumulative Proportions.

β-Error and Its Relation to the Base Scale

So far we have found it unnecessary to make any comments concerning the base scale which supposedly underlies the rating scale except to hypothesize its existence.

Necessary and sufficient conditions for the existence of an *interval* base scale underlying a successive interval rating scale are: (*i*) There must (potentially) exist a numbering of all the potentially infinite number of events classifiable by the rating scale such that there is no overlap, for any two categories of the rating scale, of the ranges of the numbers corresponding to the events falling within each category. (*ii*) The positions of the ranges corresponding to the various categories must be in the same ordinal relation as are the categories. (*iii*) For any set of events so numbered, there must exist some interpretation of (*a*) the ordinal relations among the numbers

assigned to members of the set and of (*b*) the ratio of the difference between the numbers of any pair of the set to the difference between the numbers of any other pair, in terms of some properties of the events of the set. (If the base scale provides interpretations for additional properties of the numbers assigned to events, it may become a *ratio* or even an *absolute* scale.) There may be many different numberings satisfying conditions (*i*) and (*ii*), and many different interpretations in accordance with condition (*iii*). Hence there may be many different base scales underlying a given successive intervals scale. In fact, any assignment of numbers satisfying conditions (*i*) and (*ii*) is a potential base scale for the rating scale since we can never know for certain that there exists no interpretation of a numbering in conformance with condition (*iii*). Such potential base scales for a given rating scale need be correlated only to the extent that the values of a given event on the various potential base scales must all fall within ranges corresponding to the same rating scale category. In particular, any order-preserving transformation of a potential base scale is also a potential base scale.

The essential result of a method of successive intervals analysis is the derivation of a set of numbers corresponding to the boundaries of the intervals of the rating scale; these numbers, when paired and the ratio of differences between members of pairs taken, give the ratio of the base scale intervals corresponding to these pairs. The ratio of intervals for a potential base scale is always the same as the corresponding ratio for any *linear* transformation of that scale. However, this is not uniformly true for any other transformation. Let all potential base scales be separated into classes, any member of a given class being a linear transformation of any other member of that class. These classes are, in general, characterized by different values for the ratio of two intervals corresponding to two pairs of points on the rating scale; the classes of potential base scales most closely approximated by the scale computed through the successive intervals technique will be those classes whose ratios for interval widths are most similar to the ratios displayed by the computed scale. That is, the classes of potential base scales most closely approximated by the computed scale are the classes of potential base scales minimizing the ξ_{ik} .

Since the exact magnitudes of the ξ_{ik} are unknown in applications of the method of successive intervals, it is impossible to determine the class of potential base scales most closely approximated in a specific instance. The classes of potential base scales *most likely* to minimize the ξ_{ik} , however, are those which minimize the *expected* values of the ξ_{ik} . Now, the data of a specific successive intervals analysis are obtained by sampling of two kinds: a sample of size N from possible distributions over the rating scale, and a sample of n_i individuals from each distribution i ($i = 1, 2, \cdots, N$). The expected value of ξ_{ik} for a specific sample of distributions is given by (25). But the terms α_{ik} , $(\beta_i' - \beta_k')$, β_i'' , β_k'' are dependent upon the specific sample of

distributions chosen; β_i'' and β_k'', by definition, have an expected value of 0, while both α_{jk} and $(\beta_j' - \beta_k')$ are determined essentially by differences of the form $X_j r_{y_j z_j} - X_k r_{y_k z_k}$, where X, y, and z are various specified properties of the N distributions. These differences should be negative as often as positive, so that the expected values of α_{jk} and $(\beta_j' - \beta_k')$ should be 0. Thus, the expected value of ξ_{jk} is approximately $\mu_{\omega_j} - \mu_{\omega_k}$, and hence the classes of potential base scales most closely approximated by the expected computed scale are those base scales showing the smallest differences among the μ_{ω_j}, μ_{ω_k}, $\cdots$ for the various intervals j, k, $\cdots$ of the scale. This important conclusion may be rephrased as: *the classes of potential base scales expected to be most closely approximated by the method of successive intervals are the classes for which the average coefficient of error (for the estimation of interval widths under assumptions of normality) is most nearly the same for all intervals of the rating scale.*

In particular, if, as implicitly assumed by previous psychometric analyses wherein the base scale remained unidentified, there exists a class of potential base scales which simultaneously normalize all distributions, then $\mu_\omega = 0$ for all intervals of these scales; there is no class of base scales more closely approximated by the expected computed scale.

Thus, we see that there is no single answer to the question of the magnitude of error involved in the approximation of an unidentified base scale by the method of successive intervals; the magnitude of error is relative to that base scale for which the computed scale is considered an approximation. If we wish, however, we may define the base scale to be approximated as that scale which simultaneously equalizes the μ_ω for all intervals. A class of such scales can always be found, and further, the set of all such classes includes all base scales which simultaneously normalize all distributions over the rating scale if such scales exist. If the base scale is so defined, then from (25)

$$\mu_{\xi_{jk}} = \alpha_{jk} + (\beta_j' - \beta_k') + (\beta_j'' - \beta_k''). \tag{27}$$

Only when the measuring distributions are extraordinarily non-normal are any of the terms on the right side of (27) of expected magnitude greater than 10^{-2}, and thus ξ_{jk} has an expected order of magnitude of no greater than 10^{-2}. This, in conjunction with (26), shows that if the sample sizes of the distributions have been taken sufficiently large (say, large enough to make σ_γ on the order of 10^{-2}), then *the extent to which interval ratios computed by the method of successive intervals diverge from the corresponding theoretically "true" values should not exceed* 10 *per cent of the latter, and may be much smaller if the experimental study has been well designed.*

Conclusions

Abstracting the essentials of the foregoing analysis, three major points are of significance—the first, a contribution to the computation technique of the method of successive intervals; the second, an evaluation of the validity

of the method; and the third, the significance of the method for the basic methodology of psychophysical measurement.

The contribution to computational technique is given by equations (13) and (14); it involves the computation of weights for the estimates of a given interval width so as to minimize the sampling errors for the composite estimate of the interval width. Except for the more exacting studies, however, or unless suitable tables have been obtained, the improvement of this exact method of weighing over the more rough and ready techniques now in use will scarcely be worth the extra computational labor. Of greater potential application in the design of empirical studies is the determination of the, relations among width of interval, the number of measuring distributions and their sample sizes for the maintenance of a fixed level of freedom from sampling error.

The validity and reliability of the method of successive intervals do not depend upon normality of distributions or equality of their variances. The reliability, as attested by (26), may be made as high as desired. If the base scale is suitably defined (i.e., defined so as to equalize, for the various intervals, the error due to estimation of interval width from a table of the normal probability integral) and if the reliability is made sufficiently high, then the validity, as implied by (27), is so high as to lead to an expected coefficient of error for relative interval widths of no more than a few parts in a hundred. Further, this validity is in reference to the theoretical values of the interval ratios. It is thus an *absolute* validity in contrast to past validation of psychophysical scaling techniques, where validation is attempted only in terms of internal consistency or consistency among different techniques purported to compute the same base scale. It would appear, then, that until similar analyses can be constructed for other psychophysical scaling techniques, the method of successive intervals should be accepted as the basic standard against which other techniques are to be validated.

Finally, and probably most important of all, we consider the implications of this analysis for the methodology of psychophysical measurement. It has been shown that it is unnecessary for psychophysical measurement (or for that matter, for any form of measurement) to assume any specific form of distributions over a measuring scale. The only assumption required is that certain properties of the measurements obtained by the measuring technique have some potential interpretative significance. The major premise of psychometric scaling in the past has been that if (*a*) a scale can be obtained which normalizes the distributions over it, then (*b*) that scale, or another very similar to it, has interpretive significance as an interval scale. We may now replace this premise with another: if (*a'*) a scale can be obtained which equalizes, for all intervals, the average coefficient of error for the approximation of interval width by the distance which normal distributions of equal standard deviations would span between corresponding percentiles,

then (*b*) that scale, or another very similar to it, has interpretive significance as an interval scale. The latter premise is both weaker and stronger than the former: weaker in that a scale satisfying (*a'*) can always be found, and such a scale also satisfies (*a*) when scales satisfying (*a*) exist; stronger in that the latter premise demands a meaningful scale to underlie every psychophysical measuring technique, whereas the former demands such a meaningful understructure only if a psychometric scale can be found to normalize simultaneously all distributions over it. Actually, the (*b*) clause of these premises is not so strong as it might appear. In a certain sense, the mere act of defining a scale in terms of the distributions over it imparts a meaning to the scale values so defined. Essentially, what our present analysis has shown is that it is always possible to give a distributional definition to a base scale which simultaneously normalizes all distributions regardless of whether or not a scale exists.

Since interpretation of psychometric scales has been sought in actual practice, regardless of whether simultaneous normalization could be realized, it is essential, if psychometric custom now current is to be justified, that a way be found to define psychometric scales in terms of properties other than such normalization. It is our belief that such justification has now been furnished.

REFERENCES

1. Attneave, F. A. A method of graded dichotomies for the scaling of judgments. *Psychol. Rev.*, 1949, **56**, 334–340.
2. Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. appl. Psychol.*, 1952, **36**, 118–122.
3. Edwards, A. L. and Thurstone, L. L. An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 1952, **17**, 169–180.
4. Garner, W. R. and Hake, H. W. The amount of information in absolute judgments. *Psychol. Rev.*, 1951, **58**, 446–459.
5. Guilford, J. P. The computation of psychological values from judgments in absolute categories. *J. exp. Psychol.*, 1938, **22**, 32–42.
6. Guilford, J. P. Psychometric methods, 2nd ed. New York: McGraw-Hill, 1954 (pp. 223–262).
7. Mosier, C. I. A modification of the method of successive intervals. *Psychometrika*, 1940, **6**, 101–107.
8. Saffir, M. A. A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 1937, **2**, 179–198.
9. Stevens, S. S. Mathematics, measurements, and psychophysics. In S. S. Stevens (Ed.): Handbook of experimental psychology. New York: Wiley, 1951.
10. Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.*, 1927, **34**, 273–286.

Manuscript received 3/10/54

Revised manuscript received 3/18/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

RELATIONSHIPS BETWEEN TWO SYSTEMS OF FACTOR ANALYSIS

C. W. HARRIS
UNIVERSITY OF WISCONSIN

Considering only population values, it is shown that the complete set of factors of a correlation matrix with units in the diagonal cells may be transformed into the factors derived by factoring these correlations with communalities in the diagonal cells. When the correlations are regarded as observed values, the common factors derived as a transformation of the complete set of factors of the correlation matrix with units in the diagonal cells satisfy Lawley's requirement for a maximum likelihood solution and are a first approximation to Rao's canonical factors.

One of the distinctions in factor analysis that may be made, when viewed at the procedural level, is the distinction between choosing to factor R_1 , the matrix of the intercorrelations of the variables with units in the diagonal cells, and choosing to factor R, the matrix with communalities in the diagonal cells. The purpose of this paper is to develop the transformation which relates the factors F_1 of R_1 to the factors F_2 of R. In order to develop this relationship, it will be assumed first that the elements of R_1 and of R have been determined without error, i.e., that the correlations and the communalities are *population* values. Guttman (**2**) has discussed conditions that are necessary for common-factor, or communality, solutions. Considering only population values, this paper shows that if a communality solution exists, it is simply a projective transformation of the set of complete factors of R_1 .

The problem of estimation will next be considered. It will be shown that the transformation developed here when applied to *observed* data yields common factors that satisfy Lawley's (**5**, **6**) requirements for a maximum likelihood solution and are a first approximation to factors derived by Rao's (**7**) canonical factor analysis. In developing these points, certain simplifications have been introduced for convenience. For example, only orthogonal factor solutions are employed; however, such solutions may be rotated to oblique factors if such are desired. Second, the factoring of variances and covariances, rather than R_1 , is considered only incidentally in connection with Rao's canonical factor analysis. Third, it is generally assumed that R_1 , and consequently F_1 , is nonsingular. This is a realistic assumption. If necessary it can be abandoned, provided we make certain modifications in the algebra; these modifications do not negate the generalizations given here.

The distinction between choosing to factor R_1 and choosing to factor R is of considerable importance theoretically. In the former case the factors

are conceived to be located within the space defined by the variables viewed as vectors, whereas in the latter case both the common and the unique factors are conceived to be located in a space distinct from the space defined by the variables viewed as vectors. Consequently, factoring R_1 is equivalent to developing both the factor coefficients and the factor scores by appropriate linear operations on Z, where $ZZ' = R_1$; whereas the communality principle implies that a projection of Z, say ZA, is the matrix that is operated on to develop both the factor coefficients and the factor scores. Cf. (3). Since A is unknown, it follows that for the communality principle the factor scores cannot be computed, even though the elements of R are determined without error, but must be estimated.

It is well known (1, 4) that we may define an arbitrary orthogonal factor solution in terms of linear operations on a rectangular matrix of order n by N. Let Y be any n by N matrix, $n < N$. Then we may resolve Y into a product of two matrices, $Y = FS$, such that SS' is an identity matrix of the appropriate order. It is conventional to call F the factor matrix and S the matrix of factor scores. Various factoring methods, such as the centroid, the diagonal, or the principal-axis methods, can be identified in terms of rules for choosing the linear operators.

Now let us examine the relationship between factors extracted from R_1 and those extracted from R. If both Z and ZA are given, we may write

$$Z = F_1 S_1 , \tag{1}$$

by which is implied that the matrix Z has been factored completely, usually with as many factors as variables; and

$$ZA = F_2 S_2 , \tag{2}$$

with the implication that ZA has been factored completely, yielding $m \leq n$ factors. It has been shown (3) that $S_2' S_2$ defines A, which is the symmetric idempotent matrix that achieves the projection of Z that is required by the communality principle. Therefore, since $S_2 S_2' = I$ of order m, we may write

$$ZS_2' S_2 = F_2 S_2 ,$$

and

$$ZS_2' = F_2 .$$

Then the desired relationship is given by

$$F_2 = F_1 (S_1 S_2') . \tag{3}$$

The matrix in parentheses in (3) will ordinarily be singular; if so, it is not possible to solve (3) to write F_1 as a transformation of F_2 .

The matrix in parentheses in (3) is the transformation that relates the factors derived by the communality principle to those derived from the complete factoring of R_1 . That this is the relationship to be expected on other grounds is readily seen. Any matrix of unit-length factor scores, S, may be viewed as a matrix of direction cosines that gives the location of the factors with respect to the N person vectors, i.e., they locate the factors in the person space. With two sets of orthogonal factors located in the same person space, the transformation of the inner products of the variables with one set of factors to the inner products of the variables with the other set of factors is given by the correlations between the two sets of factors. It should be noted, however, that this is not a conventional orthogonal rotation; for example, the sums of squares of the entries in any column of the transformation need not be unity.

A solution for the transformation $T = (S_1 S_2')$ may be obtained. For example,

$$(F_1'F_1)^{-1}F_1'F_2 = T$$

follows from (3). This is a true equation and not merely a least squares approximation under the conditions described here. This is to say that the symmetric idempotent matrix generated from F_1 is a unit for multiplication of F_2 . If R_1 is nonsingular, then F_1 also is nonsingular, and so the solution for T might be written

$$(F_1)^{-1}F_2 = T.$$

Another means of solving for T follows from the requirement that

$$F_2F_2' = R = F_1TT'F_1' = R_1 - U^2 = F_1F_1' - U^2,$$

where U^2 designates the matrix of unique variances. Assuming R_1 nonsingular,

$$TT' = I - (F_1)^{-1}U^2(F_1')^{-1}. \tag{4}$$

A solution for T (or some orthogonal rotation of T) is then given by factoring the matrix on the right of (4). At this point a solution for T is merely of theoretical interest, since if population values of correlations and communalities were known, the straightforward approach to determining the common factors would be to factor R.

Estimation of Common Factors

An important statistical problem faced by the factor analyst who wishes to employ the communality principle is that of estimation. Lawley (**5**, **6**) has presented a maximum likelihood solution that has certain optimum characteristics from the statistical point of view. He also presents a test of significance for common factors derived by his iterative procedure. Rippe

(8) has extended the test of significance to factors derived by any method. Lawley's requirement on the factors F_2 is, in the notation employed here,

$$DF_2' = F_2'(U^{-2}R_1 - I), \tag{5}$$

where D is a diagonal matrix. Regard $R_1 = F_1F_1'$ as observed values. Lawley's development shows that given m common factors with U^2 as estimates of the unique variances of the variables, values of F_2 that satisfy (5) generate a reproduced matrix R^* such that a test of significance of the residual matrix $R_1 - R^*$ may be made.

Modify (5) by substituting $F_1T = F_2$, and $R_1 = F_1F_1'$; then with F_1 non-singular,

$$DT' = T'(F_1'U^{-2}F_1 - I). \tag{6}$$

This states that to satisfy Lawley's requirement, each column of T must be proportional to a unit-length characteristic vector of the matrix $(F_1'U^{-2}F_1 - I)$. From (4) it is evident that a satisfactory solution for T is to choose each column as proportional to a unit-length characteristic vector of the matrix on the right of (4), i.e., to define T by a principal-axis factoring of this matrix. But this is merely restating the requirement of (6), since the characteristic vectors of $(F_1'U^{-2}F_1 - I)$ and of $[I - (F_1)^{-1}U^2(F_1')^{-1}]$ necessarily are the same. We have therefore shown that the solution $F_2 = F_1T$ satisfies Lawley's requirement when T is defined by a principal-axis factoring of the matrix on the right of (4).

Now it is at least intuitively evident that one or more columns of T might be null or consist of imaginary numbers. Let us adopt a rule that refuses to admit such columns in any solution of F_2 . This implies that the number of common factors must be considered, as well as the estimates of unique variance. If at least one column of T is admissible, then the resulting factor or factors generate a reproduced matrix such that the residuals may be tested for significance by Lawley's procedure. If the residuals are not significant, then an upper bound to the number of common factors has been determined. If they are significant, new trial values of U^2 must be employed, and a new determination of the number of admissible columns of T and of the significance of the residuals must be made.

A relatively new attack on the problem of estimation of common factors is given by Rao's (7) canonical factor analysis, which is one of many possible maximum likelihood solutions. Rao derives his basis of estimation by requiring the correlation between a linear combination of ZA and a linear combination of Z to be a maximum. He calls this canonical factor analysis because of its connection with canonical correlation theory. His estimation procedure modifies initial trial values of U^2 by an iteration process, under the restriction of a given number of common factors. A test of significance for a least number

of factors is provided. Rao also shows that canonical factor loadings derived from correlations are proportional to those derived from covariances, with the constants of proportionality given by the sample standard deviations of the variables. Therefore, in showing that the solution $F_1T = F_2$ is a first approximation to Rao's canonical factors, we imply that the solution based on variances and covariances, instead of R_1, also is a first approximation to canonical factors.

Rao's procedure requires that we select the non-zero elements of U^2 to satisfy

$$1/u_i^2 = (\lambda_1 - 1)a_{i1}^2 + (\lambda_2 - 1)a_{i2}^2 + \cdots + (\lambda_m - 1)a_{im}^2 + 1, \qquad (7)$$

where each λ_j is a latent root of the matrix $U^{-1}F_1F_1'U^{-1}$, and each a_{ij} the appropriate element of the unit-length characteristic vectors of the same matrix. Now these roots and vectors are connected with the roots and vectors of TT' in a regular manner. This connection is derived from the fact that the roots of $U^{-1}RU^{-1} = (U^{-1}R_1U^{-1} - I)$ are all one less than the roots of $U^{-1}R_1U^{-1}$, and the characteristic vectors of $U^{-1}RU^{-1}$ and $U^{-1}R_1U^{-1}$ are identical. Let L^2 designate the roots of $U^{-1}F_1F_1'U^{-1}$; then the roots of TT' are $(I - L^{-2})$. Conversely, let the positive roots of TT' be designated by D_m^2. Then $(I - D_m^2)^{-1}$ yields the m roots in L^2 that are each greater than unity; call this matrix L_m^2. Let Q designate the unit-length rows of characteristic vectors of $U^{-1}F_1F_1'U^{-1}$. Then the m unit-length rows of characteristic vectors of TT' coresponding to the positive roots D_m^2 are given by $L_m^{-1}QU^{-1}F_1$. It is now evident that $P'D_m$, defining m admissible columns of T, yields by the above transformations the required values for substitution in (7). Apparently, then, for any specified U^2 we may characterize the solution $F_1T = F_2$ as a first approximation to Rao's canonical factors. This is verified by noting that Rao defines canonical factors, at any stage of approximation, by

$$F_2 = UQ'(L_m^2 - I)^{1/2}. \qquad (8)$$

Substituting for P' and D_m their expressions in terms of roots and vectors of $U^{-1}F_1F_1'U^{-1}$ and then simplifying gives

$$F_1T = F_1P'D_m = UQ'(L_m^2 - I)^{1/2},$$

which is identical with (8).

This demonstration of the connection of T with Rao's procedure is best characterized as making explicit an alternate path to canonical factors. It seems practically certain that Rao recognized the existence of this alternate path and rejected it for a very practical reason, namely, that his calculation routine is less laborious than one based on finding F_2 by way of F_1.

REFERENCES

1. Guttman, Louis. General theory and methods for matrix factoring. *Psychometrika*, 1944, **9**, 1–16.
2. Guttman, Louis. Some necessary conditions for common-factor analysis. *Psychometrika*, 1954, **19**, 149–61.
3. Harris, Chester W. Separation of data as a principle in factor analysis. *Psychometrika*, 1955, **20**, 23–28.
4. Holzinger, Karl J. Factoring test scores and implications for the method of averages. *Psychometrika*, 1944, **9**, 155–67.
5. Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. *Proc. Royal Soc. Edinburgh*, 1940, **60**, 64–82.
6. Lawley, D. N. Further investigations in factor estimation. *Proc. Royal Soc. Edinburgh*, 1941, **61**, 176–85.
7. Rao, C. R. Estimation and tests of significance in factor analysis. *Psychometrika*, 1955, **20**, 93–111.
8. Rippe, Dayle D. Application of a large sampling criterion to some sampling problems in factor analysis. *Psychometrika*, 1953, **18**, 191–205.

Manuscript received 10/6/54

Revised manuscript received 4/4/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

TESTING MESSAGE DIFFUSION IN HARMONIC LOGISTIC CURVES*

STUART CARTER DODD
UNIVERSITY OF WASHINGTON

The growth of a population of knowers of a message was studied to test a human interactance hypothesis. The conditions investigated involved *people interacting in time*, with the population pairing off randomly (i.e., determined by many, small, different influences) and transferring an attribute (i.e., an all-or-none act) at either a steady rate or a waning rate, subsequent to the originating stimulus. The mathematical expressions for these pre-conditions were the differential equations for the linear logistic for steady acting and the harmonic logistic for waning acting. Variant forms of these curves were developed. Two exploratory experiments, or pretests, comprised launching a coffee slogan in a town and imitating a badge wearer in a boys' camp. Since the activity rate waned harmonically in both cases, the harmonic logistic fit best in both the town and the camp as expected by the hypothesis.

I. *The Need for Waning Interaction Models*†

The Washington Public Opinion Laboratory is studying the principles of human interaction in the form of diffusing messages from person to person, using questions such as: How fast will a message spread (under specified conditions)? How far? How fully and how faultlessly? How effectively in arousing belief, retelling, and compliance? What conditions will maximize social diffusion? The principles should be stated in operational rules such as mathematical curves or other models. Each model should specify (a) the variables, (b) the social preconditions expressed as mathematical assumptions, (c) the consequent curve or formula which expresses their expected joint functioning, and (d) the procedures for testing the fit of the model. All this is a case of applying dimensional methods of analysis in the field of social physics (**1**, **2**).

This paper deals only with the questions: How fast will a message

*A paper read before Section K sponsored by the Committee on Social Physics of the AAAS Conference in Boston, December 30, 1953. This research was supported in part by the United States Air Force under Contract AF 33(038)-27522, monitored by the Human Resources Research Institute (now, Officer Education Research Laboratory, Air Force Personnel and Training Research Center), Air Research and Development Command, Maxwell Air Force Base, Alabama. Permission is granted for reproduction, translation, publication and distribution in part and in whole by or for the United States Government.

†For the Air Force, the project seeks to improve the leaflet weapon in psychological warfare. The Air Force needs to know how to maximize the desired effects of the leaflets they will drop (and have dropped by the millions) on enemy or captive populations or on our own population in an emergency. For a published description of this project and some of its findings, see (**4–9**); for the interactance hypothesis and dimensional theory, a special case of which is presented in this paper, see (**1–3**).

spread? What will be its growth curve relating people-knowing-the-message to time elapsed when all other factors are constant? It is limited to two cases where an all-or-none act spreads through a homogeneous population whose acting is either (a) steady in time or (b) waning since the stimulus that started it. Pretests indicate that the activity of retelling the message tends to wane inversely with time. Since these observations seem to call for either linear logistic curves or harmonically waning logistic curves, this paper is a report on developing and testing such models.

II. *The Social Preconditions*

The social preconditions that are assumed in the simplest case may be stated in terms of three factors (at their first powers): (1) a human *population*, P; (2) an *activity*, A; (3) a period of *time*, T.

The relation at issue is that of the population-showing-the-activity to time (i.e., $PA^0 \alpha T$). How does the number of knowers of the message grow with the time elapsed since the message started?

The preconditions should also be stated in terms of these three factors only. This proves possible by stating them in terms of maximizing a proportion of the *zero*-th statistical moment and minimizing second moments or powers of these factors. The preconditions assumed may be loosely stated here (and operationally further defined with testable indices below) as follows:

1. The population interacts randomly. Everyone has an equal opportunity to interact—a meeting being determined by *many*, *small*, and *uncorrelated* influences. It seems likely that this condition may be approximated in a population that is sufficiently large and homogeneous for probability principles to work out smoothly.
2. The time is sufficiently long to observe most of the growth or diffusing of the activity, i.e., from two hours up to three days.
3. The activity is a novel all-or-none act of any person upon any other person. The first occurrence of a one-way all-or-none activity is chosen at first for simplicity in studying interaction, here the retelling of a message.
4. The activity rate or "potency rate," defined as acts-per-actor-per-period, is either: (a) steady from period to period—the linear case; or (b) waning with the time elapsed since the start—the harmonic case.

The acting may be expected to be steady when the stimulation is steady for everyone, as in the case of some dated goals ahead. It may be expected to wane with time under one or more of at least three possible causal mechanisms which might be labeled (a) *overlaying*, (b) *dropping out*, and (c) *rebuffing*.

Whenever an event is the unrepeated stimulus and it becomes overlaid with other somewhat equivalent events or interests in each of the n succeeding time periods, the first event will then tend to be reduced by a factor of $1/n$ in the public's attention. The public's responding to such a punctiform stimulus will then decrease inversely with the time since it happened.

The acting may also be expected to wane with time as a *drop-out* effect whenever there are individual differences in the speed and/or output of activity. If the output is fairly constant, faster actors will finish and drop out, leaving fewer and slower actors in the latter periods. Similarly, if output varies but speed is fairly constant from person to person, then those of low output will finish early and drop out, leaving progessively fewer actors. If both output and speed vary, the dropping out is accelerated and the average activity rate must wane as time goes on, the form of this curve of waning activity depending on the distribution of the individual differences.

The interacting may wane with time due to the tellers getting more and more rebuffs. As more and more people hear the message, they will forestall its teller. He will stop trying to tell the message if he gets rebuffed often enough. This explanation, like that above, will result in a slackened pace of telling and of individuals ceasing to tell or *dropping out.* But the cause in rebuffing is social, whereas the cause in the other case is more specific to the individual.

The two major conditions of human interaction here (namely, that the population meet randomly and that the activity rate be specified), are highly general. They transcend any local culture or transient situation and apply to any all-or-none novel behavior. Personal, situational, and cultural conditions may affect the numerical size of the activity rate of a particular act (or message) in a particular population and a particular situation. But if the given activity rate be either steady or waning in a large homogeneous population, then a logistic curve of growth is hypothesized to be the necessary consequence.

III. *The Mathematical Derivation of Models Matching the Preconditions*

A. The Linear Logistic. The derivation of the linear logistic curve assumes a large population (at least over 100 and preferably over 1,000) which is divided into two proportions: p, the proportion of knowers, and q, the proportion of nonknowers, at any moment, so that $p + q = 1$. Let them mix thoroughly during a unit period. *This social interacting is mathematically represented by* an *overtelling* proportion, p^2, of knowers meeting knowers; a *first-telling* proportion, pq, of knowers meeting nonknowers, and a *nontelling* proportion, q^2, of nonknowers meeting nonknowers. Assuming independent probabilities, pq is the probability of a meeting in which the message can be spread. Let k represent the conditional probability of an actual telling. Thus, pq is the probability of a knower and nonknower meeting, and k is the probability of telling-if-met. k is observable from the "activity rate" or "potency rate" defined as the hearers-per-teller-per-period. The product, kpq, is then the net probability of a first telling during a unit period. This is the expected rate of growth of the message. Written as the differential

equation for an increment of diffusion in an increment of time,

$$\frac{dp}{dt} = kpq. \tag{1}$$

k is assumed to be constant from period to period as p increases and q decreases. Integrating (1) over time gives the linear logistic, a symmetric S-shaped growth curve,

$$p_t = \frac{1}{1 + \frac{q_0}{p_0} e^{-kt}} \quad \text{or} \quad \frac{p_t}{q_t} = \frac{p_0}{q_0} e^{kt}, \tag{2}$$

where p_0 and q_0 denote the knowers and nonknowers at the start, and where $k/4$ is the slope at mid-date and mid-diffusion where the slope is maximal. Thus k shows the general steepness of the curve or speed of diffusion in a general way.

This simple logistic may be generalized in many ways, one of which is to substitute a function of time, $f(t)$, for the constant k and rewrite (1) as

$$\frac{dp}{dt} = f(t)pq. \tag{3}$$

This function is k_0/t in the harmonic logistic equation below. The cumulative "augmented" logistic growth curve then is

$$p = \frac{1}{1 + \frac{q_0}{p_0} \exp\left[-\int_0^t f(x)\, dx\right]}. \tag{4}$$

A quadratic exponent giving a cubic logistic fits some of our data better than a linear logistic, but it requires four parameters instead of two, and parameters with no social interpretation at present. Still further generalizing, p and q may be replaced by integratable functions

$$\frac{dp}{dt} = f_1(t)f_2(p)f_3(q). \tag{5}$$

The subscripts here denote different functions. The dimensional family of these functions uses integral exponents, positive or negative or zero, to specify particular functions which describe many important social situations.

The linear logistic may be written in discrete form as

$$p_{t+1} = p_t + kp_tq_t\ , \tag{6}$$

where p_t is the cumulated proportion of knowers at time t, q_t the nonknowers, and $t + 1$ the next unit period.

This may be rewritten in terms of each successive proportion's being equal to the mean plus the (weighted) variance of the attribute in the pre-

ceding period, i.e.,

$$M_{t+1} = M_t + kV_t \,. \tag{7}$$

For a quick trial-fitting by plotting the data on semi-logarithmic paper, the rectified logistic is convenient:

$$\ln p/q = kt + \ln p_0/q_0 \,. \tag{8}$$

If such a plot is linear, the slope, k, is the "potency" parameter.

The logistic curve is factorable, since it is a simple product of the waxing exponential growth curve,

$$p_t = p_0 e^{k_1 t}, \tag{9}$$

times the waning exponential growth curve,

$$1/q_t = 1/q_0 e^{k_2 t} \qquad (\text{or } q_t = q_0 e^{-k_2 t}). \tag{10}$$

The right-hand side of equation (3) is their product when $k = k_1 + k_2$ (see Figure 1).

The harmonic logistic is similarly factorable into the two harmonic exponential curves. All of these are factorable in both their differential equation and their integrated forms.

A special variant form of the logistic becomes the Gompertz or "simplex" growth curve. The cumulated discrete linear logistic when $k = 1$ is also a special case of the Gompertz curve with a growth rate of 2. Thus (6) can be rearranged as

$$1 - q_{t+1} = p_t + p_t - p_t^2$$

or

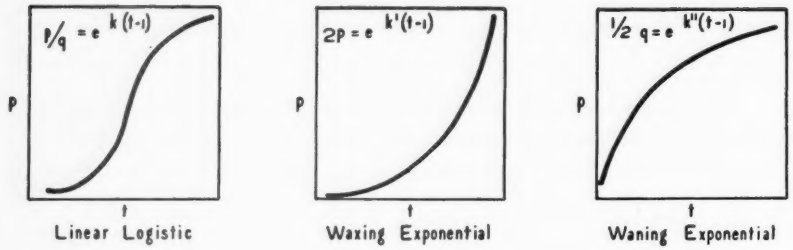
$$q_{t+1} = 1 - 2p_t + p_t^2 = (1 - p_t)^2 = q_t^2 \,.$$

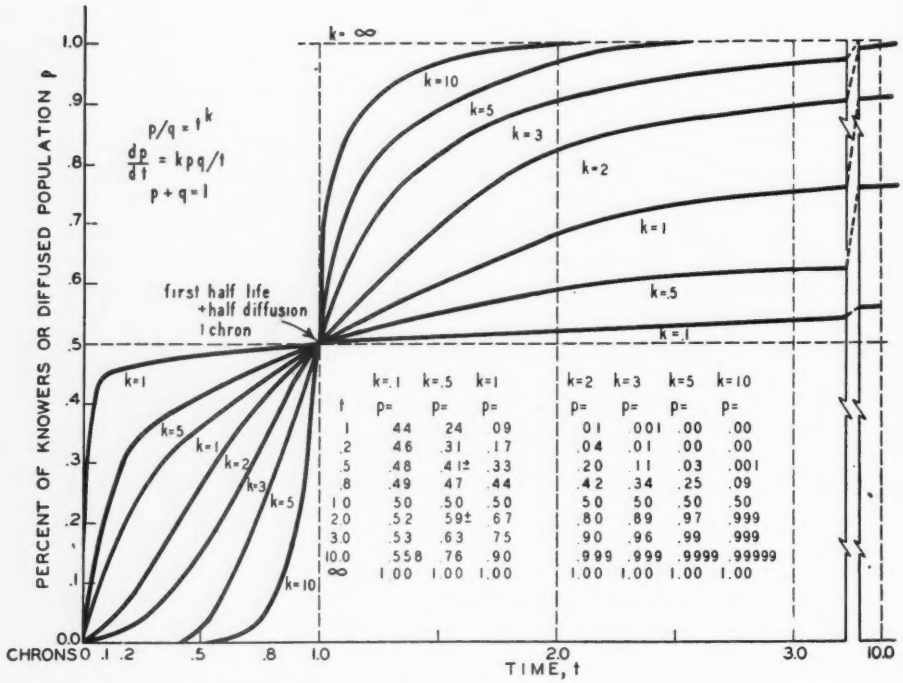
Starting with q_0 and squaring in each of the t successive periods gives

$$q_t = q_0^{2^t}. \tag{11}$$

This describes the decreasing of nonknowers, since q is less than unity and the exponent is greater than unity. If there are 99 per cent nonknowers at the start (q_0), it takes some three unit periods to shrink them to less than one per cent, i.e., $(.99)^{2^{3+}} < .01$.

B. The Harmonic Logistic. Next, if the activity rate k is not constant but is observed to reduce toward zero in time, the simplest form of descriptive curve is a harmonic series or inverse integers from one on up (or a hyperbola for continuous data). A straight line is simpler but cuts the axis and becomes meaningless when negative. The harmonic curve is preferable, since it fades away asymptotically. Its cumulative form is in terms of natural logarithms. It has many social applications. One such is Zipf's size-rank rule and the





The Harmonic Logistic Growth Curves in Chron Units

FIGURE 1

hypothesized least effort principle. In cumulative form, it can often be interpreted as an extension to new fields of the Weber-Fechner relation. Such theoretical considerations, together with the fact that a harmonic curve has fitted the time series of activity rates closely in several pretests, determined its exploration here.

The harmonically waning activity rate in the logistic equation for interacting is specified by

$$\frac{dp}{dt} = k_0 pq/(t + 1), \tag{12}$$

where k_0 is the potency rate at time 0 and $k_0/(t + 1)$ is the potency rate at any time t. The $t + 1$ means choosing $+ 1$ as the time origin to avoid having the growth rate become infinite at $t_0 = 0$, as it would in the simpler form, $dp/dt = kpq/t$. (With the simpler form, one can use the convention that growth starts at $t = 1$, not at $t = 0$, to avoid an infinite growth rate.) This integrates to the cumulative growth curve,

$$p_t = \frac{1}{1 + \frac{q_0}{p_0}(t + 1)^{-k_0}} \quad \text{or} \quad \frac{p_t}{q_t} = \frac{p_0}{q_0}(t + 1)^{k_0}. \tag{13}$$

Here p_t denotes the cumulative proportion of knowers at time t and the zero subscript denotes the starting moment (see Figure 1).

C. Units, Range, and Inflection Points. In order to standardize the abscissa units of time in the asymmetric harmonic logistic curve and so make different growth curves more comparable, the first "half life" may be taken as a standard unit. This "chron," as it may be called, is the time from *no* diffusion to *half* diffusion, from zero per cent of knowers, p, to 50 per cent of knowers. (The quarter life or any other fractional life might be used instead but with loss of simplicity.) The first "half life" does not equal the second. Since this curve is asymptotic at the upper limit only, the chron, or the period for the first 50 per cent to become knowers, is definite, while the remaining period is indefinite. In terms of chrons (with the origin at the absolute zero of diffusion) (13) becomes

$$\frac{p}{q} = t^{k_0}. \tag{14}$$

The chron may be viewed as an inverse function of the potency of the message (when comparing populations of the same size), for it is a reciprocal of an activity rate. The activity rate states "acts-per-period;" the chron states a "period-per-half-the-acts," or a time for 50 per cent of the population to be diffused. The longer the chron, therefore, the more "impotent" is the message. Thus, the chron or first half life becomes another standardized measure of strength of a message in a given population and situation. But

it is not a simple inverse potency, for it is stated in terms of a *per cent* of a total population. Thus, chrons as a measure of message strength are comparable in different populations only insofar as the populations are alike in size. Different potency rates, however, are comparable from any population since they are per capita activity rates.

The reasons for using the chron unit are simplicity of the formula and flexibility of fitting. In chron units the formula drops the explicit p_0/q_0 term and can be written in simplest form as $p/q = t^k$, or as $p = qt^k$. In chron units the node where all the family of curves intersect (Figure 1) is at the half diffusion point where $p = q = .5$. This gives a wide range of shapes from convex-up to S-shape available to fit given data. In smaller time units the node occurs so far to the left, or so early in the total growth period, that the curves are convex-up for most of their range. In that case, a harmonic logistic model might *apparently* give a bad fit even though the growth was logistic and its activity rate, k, did wane harmonically. This misleading result can happen when the time continuum is subdivided into a size of time unit and measured from an origin point which are inappropriate and obscure the harmonic waning of the potency or activity rate. The chron is an "optimal" time unit in the sense that its length is so fitted to the data in hand as to develop a versatile family of curves among which a close fit to the data may be more attainable.

In practice, the range between the upper and lower limits, or asymptotes, varies with the potency rate. The time range in the linear logistic is from $+\infty$ to $-\infty$. In practice, it has to be truncated at arbitrary points (such as $p =$ one per cent and $p = 99$ per cent), since the curve approaches 100 per cent and zero per cent asymptotically as time goes to plus or minus infinity. The harmonic logistic, however, starts from absolute zero point in time (with no knowers, so $p_0 = 0$ at t_0) and goes to 100 per cent asymptotically. For some purposes it may be truncated at 10 chrons, giving a standard time range of 100 decichrons along the abscissa to match the population range of 100 percentage points along the ordinate. Figure 1 shows these values of p at 10 chrons for various potencies.

The zero point of growth in the harmonic logistic requires extrapolating to determine it accurately and in order to fit data more closely. For the starters' date, t_A , is after the absolute zero, t_0 , from which the chron should be computed. The starters come into the curve, as it were, at the date in chron units when it has grown up to the starters' percentage of knowers, p_A (on the curve fixed by one k).

A difficulty in fitting these qt curves is to determine the absolute zero point, t_0 , in time, which occurs when p, the growth, is zero, p_0 . But in practice the first observable amount of growth is when the starters, p_A , become knowers, which is later than the true zero point by a small but unknown amount. To estimate the zero date, the starters' conversion date

may be taken as a first approximation for computing the chron and then k from the slope of the points. Using this best-fit estimate of k and the starters' proportion, p_A, as p, one can then solve for t, getting a second approximation. Further approximations may be made by repeating these steps until the correction yielded is negligible. It seems likely that if the starters are few (such as less than one per cent), the error in the chron will usually be of the order of one per cent for values of k near unity. Then a single correction will be sufficient. It will always put the starters' date at some positive fraction of a chron after the estimated absolute zero date and thus will lengthen the chron and flatten the growth curve slightly.

The upper limit of the growth curve always presents a problem. Whether chrons or other time units are used, in the case of either the linear or the harmonic logistic, it is necessary to know P_s, the terminal population, in absolute numbers. The difficulty is that there may be many different interpretations of P, such as:

1. The *census population* in a diffusion area (which may be too large, since it may include undiffusable elements such as babies who cannot talk).
2. The *relevant population* in a diffusion area which is thought a priori to be diffusable but which may have an undiffusable fraction cut off from each other by unknown barriers of physical, physiological, psychological, or cultural origin.
3. The final *diffused population*, which may be unknown and have to be estimated as generally less than the relevant population. If the latter (such as "all adults") is known, it logically should yield better predictions than the larger census population. But insofar as the diffused population differs from the researcher's a priori judgment of the relevant population, the fits will be loose and the predictions poor.

We have discussed two harmonic logistic curves, namely,

$$\frac{dp}{dt} = kpq/(t + 1), \tag{15}$$

where t is expressed in some conventional time unit, and

$$\frac{dp}{dt} = kpq/t, \tag{16}$$

where t is expressed in chron or half life units. In the case of (15), we find that the shape of the curve has the following dependence upon k: For $p_0 < \frac{1}{2}$, the curve is concave downward throughout; for

$$0 < k \leq \frac{1}{1 - 2p_0}$$

and for

$$k > \frac{1}{1 - 2p_0},$$

the curve is sigmoid with an inflection point at

$$t' = \left(\frac{q_0}{p_0} \cdot \frac{k-1}{k+1}\right)^{1/k} - 1. \tag{17}$$

In the case of (16), where time is expressed in chron units, the curve is concave downward throughout for $k \leq 1$ and is sigmoid for $k > 1$, with a point of inflection at

$$t' = \left(\frac{k-1}{k+1}\right)^{1/k}. \tag{18}$$

The harmonic logistic family of curves thus has a wide range of shapes, determined by k, combined with any steepness as determined by the chron, c.

D. "Timeless" Forms of the Harmonic Equations. It proves possible to rewrite the harmonic equations above explicitly in terms of k, p, and q alone with the time factor "cancelled out." This means that the growth can be described and predicted in terms of the knowing and nonknowing proportions exclusively without knowing whether the time units are ordinal units of removes from the first teller or cardinal units of clock time. For this, substitute the t in the equations above into their respective differential equations:

Curves	*In general*		*For* $k = 1$	
Harmonic logistic	$\frac{dp}{dt} = k_0 p^{1-(1/k_0)} \cdot q^{1+(1/k_0)}$	(19)	$\frac{dp}{dt} = q^2$	(19a)
Harmonic waxing exponential	$\frac{dp}{dt} = k_1 p^{1-(1/k_1)} \cdot \left(\frac{1}{2}\right)^{1/k_1}$	(20)	$\frac{dp}{dt} = \frac{1}{2}$	(20a)
Harmonic waning exponential	$\frac{dp}{dt} = k_2 q^{1+(1/k_2)} \cdot \left(\frac{1}{2}\right)^{-1/k_2}$	(21)	$\frac{dp}{dt} = 2q^2$	(21a)

Note that the sum of exponents on the two population factors in (19) is 2 for the logistic. Dimensionally this denotes *interaction*, i.e., a group phenomenon. The sum of the exponents in (20) and (21) is one. Dimensionally this denotes *action* (or reaction), i.e., a plurel phenomenon.

The empirical or social interpretation of (19a), (20a), and (21a) in the special case when $k = 1$ has not yet been fully explored. It appears that (20a) reflects constant growth, because the curvature of the waxing

exponential is here exactly counterbalanced by the opposite curvature of the harmonic activity rate. In (21a), the harmonically waning exponential curiously has just twice the growth rate of the harmonic logistic in (19a). One possible social interpretation of the logistic (19a) having q^2 as its growth rate seems at present to be that the growth rate depends on two factors, one physical and one psychological, which here both happen to be numerically the same, namely, q. Thus, the growth rate shrinks as q shrinks when nonknowers become fewer and physically harder to find. But, in addition, a psychological factor may operate in that as the tellers get more and more rebuffs from the growing proportion of knowers, the tellers slacken their telling activity and this slackening keeps pace with q, the proportion of current nonknowers. This explanation and alternative ones must be empirically tested.

IV. *Some Experimental Testing*

In order to begin testing the foregoing theory and to develop more definite tests, a dozen preliminary experiments were made in the first year of Project Revere. Two of these will be reviewed here as tests of the harmonic logistic hypothesis. [Some of our tests of the linear logistic have been published elsewhere (**7**, **8**, **9**)]. Both were designed to test linear logistic models in clock time units. In the first, however, the activity rate, k, was found a posteriori to wane harmonically with removes, and its growth from remove to remove should therefore fit the harmonic logistic more closely. In the second set of data to be reported, the activity rates were also observed to wane, and therefore the harmonic logistic should again give a posteriori a better fit than the linear logistic.

Time measured in removes or generations of hearers is in ordinal units and is more free of diurnal and other rhythms which "overlay" the growth curves in clock or cardinal time units. Ordinal units seem apt to yield smoother curves which fit models more readily. But curves in cardinal time units are needed wherever practical prediction of growth in clock time is wanted. In both cases the reward offered as stimulus was expected to evoke steady acting throughout the whole period. Instead it produced a spurt of activity which waned steadily—perhaps because of the overlaying or drop-out, or rebuffing mechanisms noted above.

A. Coffee Slogan Diffusing in a Town. One randomly chosen housewife of every six in a village of 950 inhabitants was told a new coffee slogan by an interviewer ringing her doorbell on Monday morning. All were invited to retell it to their friends. A free pound of coffee was promised for every housewife in town who might know the slogan when every household would be canvassed later.

This message spread from person to person till 88 per cent of the housewives knew it on Wednesday's census of households, determining the pro-

portion of knowers, the ordinates of the curve, or P values in Table 1, Column 3. In this census questions of who told whom, when, and where identified the remove of each respondent and so measured the increment of new knowers at each remove of retelling and the potency of hearers-per-teller of each remove. The potency was found to wane with successive removes in an harmonic curve. Thus, the harmonic subcase of the logistic should fit these data better than the linear case.

TABLE 1

Data for Testing the Harmonic Logistic Growth Model in "C-ville"

Removes (ordinal time units)	Chron time	Observed cumulated population of message knowers	Observed increment of knowers	Expected increment of knowers by the harmonic logistic model
t	t_c	P	ΔP	$\Delta P'$
0	.84708	42	22.83%	21.46%
1	1.05876	111	37.50	39.52
2	1.27044	164	28.80	25.68
3	1.48212	178	7.61	8.93
4	1.69380	180	1.09	2.81
5	1.90548	184	2.17	0.96
Totals	1.9 "first half-life" units	184 housewives	100.00%	99.36%

1 remove = .21168 chrons

The observed potency rates for the successive removes showed a closeness of fit correlation coefficient of .99 with the best fitting simple harmonic curve ($a = k/t$). (See Table 1).

The correlation coefficient of the increments of the observed growth of message knowers with the increments in growth expected by the harmonic logistic curve ($p/q = t^k$) was also .99 (i.e., $r_{\Delta P \Delta P'} = .99$). By the z test this r is significantly different from zero and also from our arbitrary standard of close fit, namely, $r = .9$ at the 5 per cent level. (Exactly how applicable the z test is here, however, is unknown since the variate p, "knowing the message," was dichotomously observed and may not be normally distributed; also while the starters were a 20 per cent random sample of the households, only one town was studied, and without replication this may not be representative of other communities.)

The standard for nonrejection of a hypothesis was that (a) the closeness of fit correlation index should exceed .9 between the observed and the model-

expected data in uncumulated form, and (b) this r should be significantly different from zero at the five per cent confidence level. (This generally entails an r in cumulated data, such as is usually reported, above .99, but this test is insensitive and partly spurious since cumulating compels some correlation even in random series.)

The closeness of fit correlation of the uncumulated data to the linear logistic curve which is based on a steady activity rate was .37. This linear logistic hypothesis then was rejected. But the similar closeness of fit correlation (by a successive approximations technique in fitting) of the uncumulated data to the harmonic logistic, based on a waning activity rate, was .99. Therefore the harmonic logistic hypothesis could not be rejected.

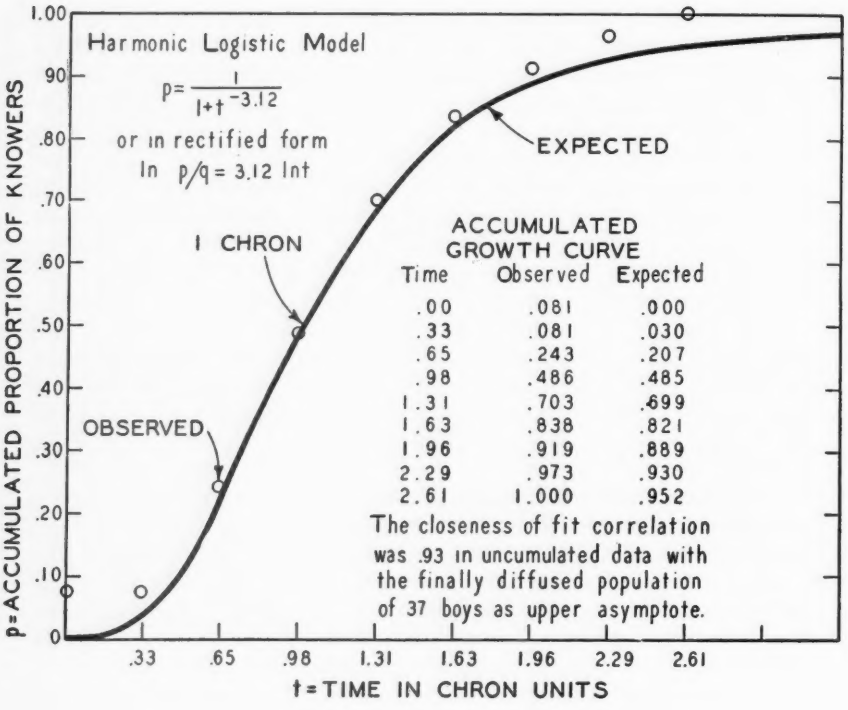
An excellent fit of one model to given data does not preclude fits as good or better by other models. In fact, the waning random net variant of the logistic, which has a constant "potency ever," fitted these uncumulated data with $r = .996$ as reported by Professor Anatol Rapoport, who developed this random net model. Also, the logistic in discrete form (6) often seems to fit our data better than the continuous curve.

B. Contagious Behavior in a Boys' Camp. Three boys in a summer camp of 42 boys were given large yellow buttons with the name of the camp and a question mark in black lettering on them. These "starters" circulated during the noon rest period and, when asked, said that they had been told a few more such buttons were obtainable at the lodge where they received their buttons. The exact time at which each boy came in to ask for a similar button was recorded and so the growth curve could be plotted. The growth rate of hearers-per-teller waned in the successive 15-minute periods during the two hours in which 39 of the boys came in for buttons. The total growth data were fitted to both linear and harmonic logistic curves (Figure 2). The closeness of fit correlations were computed on uncumulated data and were significantly different from zero at the one per cent confidence level. The correlation was .88 for the linear logistic and .93 for the harmonic logistic. The slightly better fit of the harmonic is in line with the theoretical expectations but the difference was too small and the underlying population also too small to warrant much dependence upon these findings.

For both of the tests described here, the chi square test showed that the discrepancies between model and data were not significant at the five per cent level:

	First Pretest	*Second Pretest*
Chi square	6.163	1.660
Degrees of freedom	4	6
Probability	$.90 < p < .95$	$.10 < p < .20$

We conclude that in both pretests the looseness of fit, i.e., discrepancy of model and data, was both descriptively small and statistically not significant



Diffusion of a Message in a Boys Camp Population
Fitted by a Harmonic Logistic Growth Curve in Chron Units

FIGURE 2

at the five per cent level. The discrepancies here are unimportant practically and may be due to sampling error.

The tests above were concerned with "hypotheses of form," not "hypotheses of amount." These hypotheses asserted the *form* of relation between the variables as being a linear logistic curve, etc., and did not assert the *amount* of each parameter of that curve—which could not be expected in wholly new situations.

The amounts or sizes of the parameters were determined by least squares techniques to find the best fit. Then the Pearson correlation coefficient was used to measure how closely this best fitted curve corresponded to the observed data. The technique of fitting matches the mean of the model to the mean of the data and similarly matches the two variances, leaving only

the variable discrepancy to be measured by the correlation. For this reason, the Pearson r is here almost identical with the intraclass r. The latter r is the more exacting descriptive statistic of closeness of fit, since it can approach unity only if the mean, variance, and rank order of the data agree with these moments in the model.

REFERENCES

1. Dodd, Stuart C. Dimensions of society. New York: Macmillan, 1942.
2. Dodd, Stuart C. Systematic social science. American University of Beirut, Social Science Series No. 6, 1947. (University Bookstore, Seattle, agents).
3. Dodd, Stuart C. The interactance hypothesis: a gravity model fitting physical masses and human groups. *Amer. sociol. Rev.*, 1950, **15**, 245–256.
4. Dodd, Stuart C. On all-or-none elements and mathematical models for sociologists, *Amer. sociol. Rev.*, 1952, **17**, 167–177.
5. Dodd, Stuart C. Testing message diffusion from person to person. *Public Opinion Quarterly*, 1952, **16**, 247–262.
6. Dodd, Stuart C. Testing message diffusion in C-ville. Proceedings of the Pacific Sociological Society, *Research Studies of the State College of Washington*, 1952, **20**, 83–91.
7. Dodd, Stuart C. Testing message diffusion in controlled experiments. *Amer. sociol. Rev.*, 1953, **18**, 410–416.
8. Dodd, Stuart C., and Winthrop, Henry. A dimensional theory of diffusion—an analysis, modeling and testing of one-way interaction. *Sociometry*, 1953, **16**, 160–202.
9. Dodd, Stuart C. Can the social scientist serve two masters—an answer through experimental sociology. Proceedings of the Pacific Sociological Society, *Research Studies of the State College of Washington*, 1953, **21**, 195–213.
10. Dodd, Stuart C. Diffusion is predictable. *Amer. sociol. Rev.* In press.

Manuscript received 3/26/53

Revised manuscript received 5/9/55

PSYCHOMETRIKA—VOL. 21, NO. 2
JUNE, 1956

A METHOD FOR OBTAINING AN ORDERED METRIC SCALE*

SIDNEY SIEGEL

PENNSYLVANIA STATE UNIVERSITY

A method is presented for collecting data which will yield a scale on which the entities are ranked in preference (ordinality), the distances between the entities on the scale are ranked (ordered metric), and all combinations of the distances are ranked (higher-ordered metric). The sources drawn upon are von Neumann and Morgenstern (9), and lattice theory. An empirical example is given in which a higher-ordered metric scale is derived.

If an individual is both consistent and transitive in his preferences with respect to a group of discriminable entities, it is possible by the method of paired comparisons to rank these entities in the sense of an ordinal scale of utility, e.g., $A > B > C > \cdots > N$ (read: A is preferred to B, etc.). An individual is *consistent* if he prefers the same entity of a pair whenever that pairwise comparison is presented to him. His preferences are *transitive* if when $A > B$ and $B > C$, then $A > C$. The entities involved (A, B, C, etc.) may be objects or actions. The *utility* of an entity is, roughly, the subjective value of that entity.

Such a scale, however, gives no information about the relative sizes of the differences in utility between the entities. Coombs (**2**, **3**, **4**), among others, has shown that knowledge of the magnitudes of these differences would strengthen the measurement of the psychological attribute involved and, therefore, increase the amount of information obtained from the responses made by the individual to the stimuli presented to him.

Coombs (**2**) suggests the label *ordered metric* for those scales which give not only an ordering of entities but also at least a partial ordering of the distances between the various entities. Coombs also presents a method, which he calls the *unfolding technique*, for obtaining an ordered metric scale (J scale) from a rank order preference scale (ordinal I scale).

A method is developed in this paper for obtaining an ordered metric scale of preference. This method is particularly suitable for the measurement of utility, which is a central concept in decision theory. The ordered metric

*I am grateful to Professor William L. Lepley (Department of Psychology) and Professor Jack R. Tessman (Department of Physics) for their critical reading of this paper. Paul Hurst and Robert Radlow participated in many discussions on the form of measurement discussed in this paper, and assisted in collecting data. I am also grateful to Professor T. C. Benton (Department of Mathematics) for certain source materials.

scale derived by this method yields the following information:

a. It orders the entities involved, i.e., $A > B > C > D > E \cdots$.

b. It orders the distances between the entities, i.e., say $\overline{DE} > \overline{AB} > \overline{CD} \cdots$, where $\overline{AB}$ implies the difference in utility between A and B.

c. It orders all possible combinations of contiguous distances between entities, i.e., say $\overline{AB} + \overline{BC} > \overline{BC} + \overline{CD} + \overline{DE}$, or $\overline{AB} > \overline{CD} + \overline{DE}$.

This method is not restricted by the number of entities to be scaled.

Because this method yields more than a partial ordering of the distances between the entities (see *b* and *c* above), it is suggested that this type of scaling be termed *higher-ordered metric*. Coombs (**4**) seems to mean this type of scaling by his term *ordered ordered*. However, a scale may be *ordered ordered* and still not satisfy *c* above. (The author has resisted the temptation to call the present technique *ordered ordered ordered*.)

The author has collected data which permit a higher-ordered metric scaling of preferences for up to seven entities; however, this paper will present only those data concerned with preferences among five entities. The entities employed were books, phonograph records, or money.

Sources of the Method

Von Neumann and Morgenstern (**9**, p. 17) have suggested that measurement of a person's utility in a stronger sense than ordinality could be obtained if

a. the person can always say whether he prefers one entity to another, and

b. the person can also completely order probability-combinations of the entities, i.e., combinations of entities with stated probabilities of attainment, e.g., $(B, A; p)$—read: the combination of B and A,. with probability p of getting B, and probability $(1 - p)$ of getting A.

Condition *b* requires some explanation. Suppose for a given individual $A > B > C$. This individual is given a choice between $(A, A; 1/2)$, i.e., getting A for sure, and $(B, C; 1/2)$, i.e., getting B if say head occurs on the toss of a fair coin, and getting C if tail occurs. It is clear that the individual will prefer the first alternative, which is $(A, A; 1/2)$, since $A > B > C$. In probability-combinations of entities, say $(x, y; 1/2)$, the prospect is of getting *either* x or y. The probability of getting x is .50; the probability of getting y is the remaining .50. The two alternatives in the probability-combination are mutually exclusive; the individual is absolutely certain of getting either x or y if he chooses that probability-combination.

We expect the individual to possess a clear understanding of his prefences among the entities A, B, and C (this is condition *a*), and we also expect him to prefer getting A for sure to a 50-50 combination of B and C.

Now suppose that the individual must choose between $(B, B; 1/2)$ and

$(A, C; 1/2)$. That is, he must choose between getting B for sure or getting a 50-50 chance at A or C. By making this choice, he yields new information. If he chooses the combination which gives him B for sure, his choice indicates that B is closer to A than it is to C. If he chooses $(A, C; 1/2)$, then B must be closer to C than it is to A. This is fundamentally new information because the statement $A > B > C$ told us nothing about the distances (differences) between the entities on the utility scale. Thus, the von Neumann and Morgenstern suggestions imply the possibility of measurement of utility on at least an ordered metric scale.

The second source drawn upon in higher-ordered metric measurement is lattice theory. This source is not so centrally important as the first, but it does offer a heuristic device for indicating the minimum information necessary for achieving higher-ordered metric scaling. Birkhoff (**1**, p. 6, pp. 66-72) suggests various diagrams which give a descriptive ordering of entities. Coombs (**4**, p. 4; **5**, p. 475) suggests such diagrams and puts them to use.

The lattice used here (Figure 1) not only gives a descriptive ordering of probability-combinations of entities (based on the individual's preference rankings) but also makes apparent which probability-combinations are not orderable from just a knowledge of the preference rankings. Such probability-combinations will be called *non-orderable.*

If an individual's preferences among five entities are $A > B > C > D > E$, the probability lattice is shown in Figure 1; where there is a connecting line between two probability-combinations, the higher probability-combination is preferred to the lower (**5**, p. 475). In other words, if it is true that $A > B > C > D > E$, then any two probability-combinations on the lattice that can be connected with a line which is consistently going up (or down) can be ordered, with the higher probability-combination being preferred to the lower, e.g.,

$$(A, E; \tfrac{1}{2}) > (B, E; \tfrac{1}{2})$$

$$(A, D; \tfrac{1}{2}) > (A, E; \tfrac{1}{2})$$

$$(B, D; \tfrac{1}{2}) > (D, D; \tfrac{1}{2}), \text{ etc.}$$

Simple ranking tells us nothing about the non-orderable relations, i.e., any two probability-combinations which cannot be connected by a line always going in the same direction (with respect to the horizontal-vertical dimension), e.g.,

$$(A, E; \tfrac{1}{2}) \ ? \ (B, D; \tfrac{1}{2})$$

$$(A, E; \tfrac{1}{2}) \ ? \ (B, C; \tfrac{1}{2})$$

$$(A, E; \tfrac{1}{2}) \ ? \ (B, B; \tfrac{1}{2})$$

$$(A, D; \tfrac{1}{2}) \ ? \ (C, C; \tfrac{1}{2}), \text{ etc.}$$

It is the relations between these (non-orderable) pairs of probability-combinations which contain the information necessary to change an ordinal scale to a higher-ordered metric scale.

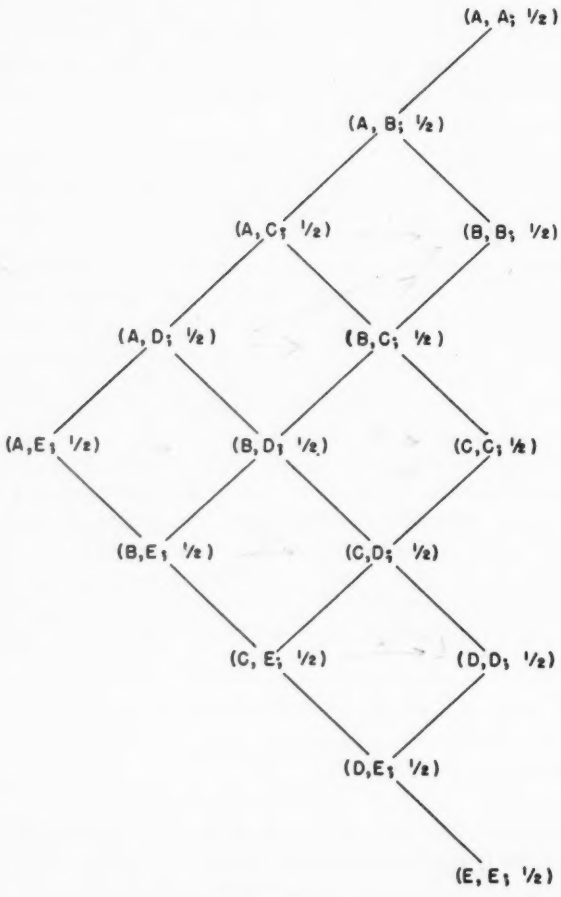


Figure 1.

A careful perusal of the lattice in Figure 1, which is based on five entities, will disclose that there are 15 pairs of probability-combinations which cannot be connected by an always-rising or always-descending line. It can be shown that if N is the number of entities to be scaled, then $\binom{N+1}{4}$ gives the number

of non-orderable pairs of probability-combinations. In the case of five entities, $\binom{6}{4} = 15$.

Not all of the non-orderable relations must be found in order to obtain the information necessary for higher-ordered metric scaling. In the present example from three to six of these relations (depending on the type of underlying ordered metric scale) are needed to achieve ordinal ordered metric scaling, and ten at most are required to achieve higher-ordered metric scaling (in which all combinations of the distances can be ordered). The fact that not all 15 pairs are needed is important. The ordering of the remaining pairs can be predicted after the minimum are used to obtain the necessary information; the success or failure of these predictions provides a check on whether higher-ordered metric scaling has in fact been achieved.

In what follows the general method of obtaining a higher-ordered metric will be outlined. Then the operational definitions necessary for implementing the general method will be given. Finally an example of the empirical use of the method will be given.

General Requirements

1. Interpret the p in the probability-combinations, e.g., $(x, y; p)$ to be *subjective* probability. [Ramsey (8) first suggested the importance of subjective probability (degree of belief) in the measurement of utility. His theory resembles that given here in many respects and antedates von Neumann and Morgenstern by more than a decade. However, the latter authors were the first to make operationally clear a method for the measurement of utility. I am indebted to Professor Donald Davidson of Stanford University for demonstrating the significance of Ramsey's work to me.]

2. Find an event for which the person's subjective probability can be experimentally determined to be one-half.

3. Require the person to rank, by the method of paired comparisons, the entities used.

4. Require the person to state his preference between each non-orderable pair of probability-combinations. (As shown above, the majority of pairs of probability-combinations may be ordered, as in the lattice in Figure 1, from a knowledge of the person's ranking of the entities. Step 4 is concerned with those pairs which cannot be ordered from this knowledge.)

5. Observe those choices which will permit the determination of an ordered metric scale.

6. Observe those choices which will permit the determination of a higher-ordered metric scale.

7. Check whether the remaining choices are consistent with the scale derived from Step 6. If all of these choices are consistent, then all of the previously non-orderable (in 4 above) pairs of probability-combinations are

consistent with each other and can be predicted from the higher-ordered metric scale derived. If Step 7 succeeds, then higher-ordered metric scaling has in fact been achieved.

Operational Definitions

Davidson, Siegel, and Suppes (6), in a study designed to measure the utility of money in the sense of an interval scale, developed an event which, for most people, has a subjective probability of one-half. Such an event is difficult to find because of the prejudices and superstitions which many people hold concerning familiar events, e.g., heads on coins, evens on dice, etc. The event used here is produced by means of specially-made dice. On three faces of the die, the nonsense syllable *ZOJ* is engraved, and on the other three faces *ZEJ* is engraved. Similar dice were made with pairs *WUH*, *XEQ* and *QUG*, *QUJ* on their faces. These syllables were selected from Glaze (7), who reports these pairs to have practically zero association value. The dice were tested with each subject; in every case the expectation of zero association was upheld, i.e., each subject was indifferent about which nonsense syllable he would bet on or which one would be the winner. The use of these dice will be discussed further in the following section.

Careful and considered choices between the probability-combinations presented to a subject were assured by "realistic" conditions. That is, when books were used as entities to be scaled, the subject was assured of getting a book or books. The identity of the book he received was a function of *all* of his choice behavior. Therefore he was highly involved in each choice. When amounts of money were used as the entities, the subject was given a sum of money (usually one dollar) at the start of the session. He gambled with that money, keeping all funds in his possession at the end of the session.

The essential device that defines operationally how the subject's choices determine ordered metric scaling is a one-person game (6) in which the subject chooses between two alternatives, each of which is a probability-combination of two outcomes. The format for each offer is:

	Alternative 1	*Alternative* 2
If event E occurs:	you get w	you get x
If not-E occurs:	you get z	you get y

The subject chooses the column; the outcome of event E determines the row. Event E might be *ZOJ*, in which case *not-E* would be *ZEJ*.

Suppose $w > x > y > z$. If the subject chooses alternative 1, then

$$(w, z; p) > (x, y; p). \tag{1}$$

If $u(w)$ is read as "the utility of w" and is interpreted as the subjective value of w, i.e., its worth to the person, then (1) can be written

$$p \cdot u(w) + (1 - p) \cdot u(z) > p \cdot u(x) + (1 - p) \cdot u(y). \tag{2}$$

If p is understood to be subjective probability, and is known to be one-half, then (2) can be written

$$u(w) + u(z) > u(x) + u(y), \tag{3}$$

and

$$u(w) - u(x) > u(y) - u(z); \tag{4}$$

i.e.,

$$\overline{wx} > \overline{yz}, \quad \text{when} \quad (w, z; p) > (x, y; p). \tag{5}$$

That is, w and x differ in utility more than y and z.

It should be noted that the distances are *directed* distances. That is, $\overline{xw}$ is the negative of $\overline{wx}$. To simplify comparisons, the convention has been adopted of always deriving the distance from the more preferred to the less preferred entity. [For example, from (3) we could get $u(z) - u(y) > u(x) - u(w)$. But since $w > x > y > z$, we multiply through by -1 to get (4) as shown.]

An Example of Higher-Ordered Metric Scaling

A graduate student in psychology served as the subject. He was shown a collection of books and was asked to choose from them the five books which he would most like to own. He was told to make this selection carefully, for he would surely receive one of the books he chose at the conclusion of the session. The books he chose, in the order of choice, were: (1) S. S. Stevens (Ed.), *Handbook of Experimental Psychology*, (2) E. G. Boring, *A History of Experimental Psychology*, (3) E. R. Hilgard, *Theories of Learning*, (4) E. R. Hilgard and D. G. Marquis, *Conditioning and Learning*, and (5) H. B. English, *A Student's Dictionary of Psychological Terms*, 4th edition.

All possible pairs of these five books were presented to the subject orally, and he was asked to state his preference as each was presented. His choices were:

Stevens > Boring ($A > B$)
Stevens > Hilgard ($A > D$)
Stevens > Hilgard and Marquis ($A > C$)
Stevens > English ($A > E$)
Boring > Hilgard ($B > D$)
Boring > Hilgard and Marquis ($B > C$)
Boring > English ($B > E$)
Hilgard and Marquis > Hilgard ($C > D$)
Hilgard > English ($D > E$)
Hilgard and Marquis > English ($C > E$)

The subject's choices were consistent and transitive; his choices would be ranked thus: $A > B > C > D > E$.

Having stated his preferences among the paired comparisons, the subject was introduced to the "game." He was allowed to become familiar with the

dice and the game by taking practice trials in which simple relations were offered, i.e., those relations which are connected by a line in the lattice [e.g., $(A, D; 1/2)$ or $(B, E; 1/2)$]. As a trial run he could choose between a 50-50 chance of getting either Stevens (A) or Hilgard (D), or a 50-50 chance of getting either Boring (B) or English (E).

The practice trials served not only to introduce the game but also to check on the consistency and transitivity of the lattice (Figure 1), because all of the choices on these simple relations should be predictable. The practice trials also served to check whether the subject's behavior was consistent with a subjective probability of one-half toward the event. This was ascertained when the subject showed indifference as to which syllable would be the winner. That is, if the *ZOJ-ZEJ* die was used and the subject was willing to make his choice between alternatives 1 and 2 (without knowing or caring which of the nonsense syllables was to be associated with which of the outcomes) it was concluded that the choice was based only on the utility of the entities involved and was independent of the particular event giving rise to the outcome.

The subject was told to consider the alternatives, announce his choice, encircle that choice on a 3×5 card, and then turn that card face down. He was told that after all sets of alternatives were presented to him, the cards would be shuffled and he would then draw one card. The alternative which he had encircled on that card would be determined by a roll of the nonsense syllable die. Thus each selection made by the subject might be the crucial one, so each one had to be made carefully.

After the practice conditions were met, critical sets of alternatives were presented to the subject. His choices are indicated by the direction of the carat. For example, the first alternative, (a), permitted the subject to choose either A or C with a 50-50 probability, or to choose getting B for sure. The direction of the carat shows that he preferred to take a 50-50 gamble on getting either A (Stevens) or C (Hilgard and Marquis) rather than to be sure of getting B (Boring). The choices were:

(a) $(A, C; \frac{1}{2}) > (B, B; \frac{1}{2})$

(b) $(B, D; \frac{1}{2}) > (C, C; \frac{1}{2})$

(c) $(D, D; \frac{1}{2}) > (C, E; \frac{1}{2})$

(d) $(C, D; \frac{1}{2}) > (B, E; \frac{1}{2})$

(e) $(B, D; \frac{1}{2}) > (A, E; \frac{1}{2})$

(f) $(D, D; \frac{1}{2}) > (A, E; \frac{1}{2})$

(g) $(A, D; \frac{1}{2}) > (B, B; \frac{1}{2})$

(h) $(A, D; \frac{1}{2}) > (B, C; \frac{1}{2})$

(i) $(B, B; \frac{1}{2}) > (A, E; \frac{1}{2})$

(j) $(C, C; \frac{1}{2}) > (B, E; \frac{1}{2})$

(k) $(B, C; \frac{1}{2}) > (A, E; \frac{1}{2})$

(l) $(A, D; \frac{1}{2}) > (C, C; \frac{1}{2})$

(m) $(D, D; \frac{1}{2}) > (B, E; \frac{1}{2})$

(n) $(C, D; \frac{1}{2}) > (A, E; \frac{1}{2})$

(o) $(C, C; \frac{1}{2}) > (A, E; \frac{1}{2})$

The inequalities (1) to (5) in the previous section show that these choices can be stated in terms of distances as:

(a')	$\overline{AB} > \overline{BC}$	(h')	$\overline{AB} > \overline{CD}$
(b')	$\overline{BC} > \overline{CD}$	(i')	$\overline{BE} > \overline{AB}$
(c')	$\overline{DE} > \overline{CD}$	(j')	$\overline{CE} > \overline{BC}$
(d')	$\overline{DE} > \overline{BC}$	(k')	$\overline{CE} > \overline{AB}$
(e')	$\overline{DE} > \overline{AB}$	(l')	$\overline{AC} > \overline{CD}$
(f')	$\overline{DE} > \overline{AD}$	(m')	$\overline{DE} > \overline{BD}$
(g')	$\overline{AB} > \overline{BD}$	(n')	$\overline{DE} > \overline{AC}$
		(o')	$\overline{CE} > \overline{AC}$

The first five relations—(a') to (e')—yield the ordered metric scale: $\overline{DE} > \overline{AB} > \overline{BC} > \overline{CD}$. When relations (f') and (g') are also considered, we have necessary and sufficient information for a higher-ordered metric scale. Relations (h') through (o') provide checks on the uniqueness of the higher-ordered metric scale derived by (a') through (g').

The ordered metric scale given by (a') through (e') may be depicted as:

$$\overline{A \qquad B \quad C \quad D \qquad E}$$

It is seen that the subject's choices (stated in distances) in relations (h') through (l') may be predicted (i.e., checked) from a knowledge of just (a') through (e'). However, in order to predict (m'), (n'), and (o'), the subject's choices in (f') and (g') must be known; knowledge of the latter provides a more powerful form of measurement.

Inasmuch as all of the choices in the relations (h') through (o') were predictable, i.e., were consistent with choices in (a') through (g'), the higher-ordered metric scale derived from choices (a') through (g') is unique and valid. This higher-ordered metric scale may be depicted as:

$$\overline{A \qquad\qquad B \quad C \quad D \qquad\qquad E}$$

We now know not only that $A > B > C > D > E$ (ordinal scale), and that $\overline{DE} > \overline{AB} > \overline{BC} > \overline{CD}$ (ordered metric scale), but also that $\overline{AE} > \overline{BE} > \overline{CE} > \overline{DE} > \overline{AD} > \overline{AC} > \overline{AB} > \overline{BD} > \overline{BC} > \overline{CD}$ (higher-ordered metric scale).

Data on a utility scale of five entities (records or books or amounts of money) have been collected from 10 subjects. Of this number, nine have been consistent; therefore it was possible to derive a higher-ordered metric scale for each of these nine. One subject showed inconsistencies in two of the relations which "should" have been predictable; therefore a unique higher-

ordered metric scale could not be constructed for him. At present, the nature of inconsistencies is being studied. One of the leads, suggested by Robert Radlow, is that inconsistencies are likely to occur in relations which involve equal-appearing intervals or combinations of intervals. The findings on the inconsistent subject seem to support this explanation. On the average, the time required to obtain a person's higher-ordered metric scale for five entities is twenty minutes.

REFERENCES

1. Birkhoff, G. Lattice theory. (Revised Ed.) New York: American Mathematical Society, Colloquium Publications, 1948, XXV.
2. Coombs, C. H. Psychological scaling without a unit of measurement. *Psychol. Rev.*, 1950, **57**, 145–158.
3. Coombs, C. H. Mathematical models in psychological scaling. *J. Amer. statist. Assn.*, 1951, **46**, 480–489.
4. Coombs, C. H. A theory of psychological scaling. Ann Arbor: Univ. of Michigan Engineering Research Institute, Bulletin, 1952, **34**.
5. Coombs, C. H. Theory and methods of social measurement. In L. Festinger and D. Katz (Eds.): Research methods in the behavioral sciences. New York: Dryden, 1953. Pp. 471–535.
6. Davidson, D., Siegel, S., and Suppes, P. Some experiments and related theory in the measurement of utility and subjective probability. Report No. 4, Stanford Value Theory Project, 1955.
7. Glaze, J. A. The association value of nonsense syllables. *J. genet. Psychol.*, 1928, **35**, 255–267.
8. Ramsey, F. P. Truth and probability. In: The foundations of mathematics and other logical essays. London: K. Paul, Trench, Trubner, 1931.
9. von Neumann, J., and Morgenstern, O. Theory of games and economic behavior. (2nd Ed.) Princeton: Princeton Univ. Press, 1947.

Manuscript received 2/7/55

Revised manuscript received 4/4/55

BOOK REVIEWS

ISADORE BLUMEN, MARVIN KOGAN, and PHILIP J. MCCARTHY. *The Industrial Mobility of Labor as a Probability Process.* Ithaca: New York State School of Industrial and Labor Relations, Cornell University, 1955, xii + 163 pp. $3.00 paper, $4.00 cloth. (*Cornell Studies in Industrial and Labor Relations*, Vol. 6.)

This is an investigation of how the theory of Markov chains can be used, or adapted, for a description of the observed movements of the labor force within the United States. The phenomenon of labor movements has no direct connection with psychometrics, but the theory of Markov chains appears to a useful tool for the construction of models of learning.* Psychometricians may therefore find interest in the book from a purely methodological point of view. In fact, probably never before has so huge a statistical material been treated numerically by probabilistic models of the Markovian variety. The analysis is made expertly and described carefully to the last detail. The problems of handling the data, of statistical estimation, and of comparing theory with observations appear thus with great clarity. The shortcomings of the method are discussed with commendable frankness.

Consider, say, nine categories of employment and add the tenth category entitled "not covered by the preceding ones." Fix an arbitrary time unit and consider the workers who at time t are in category i. At time $t + 1$ a fraction f_{ij} of them will be found in category j (where $f_{i1} + \cdots + f_{i10} = 1$). In a stable community these freqùencies will be (approximately) independent of t.

Denote by F_1 the ten by ten matrix with elements f_{ij}, and similarly by F_2, F_3, $\cdots$ the analogous matrices for an observational period of length 2, 3, $\cdots$. The simple Markov chain model assumes that the transitions from category i to other categories constitute a random choice which is in no way affected by the past history (for example, of the time spent in category i). If this were the case, the matrix F_1 should be nearly equal to the matrix P of the theoretical transition probabilities, and F_2, F_3, $\cdots$ should be close to the powers P^2, P^3, $\cdots$ of the matrix P.

The authors show how to estimate P and find that in actual practice the diagonal elements of F_2, F_3, $\cdots$ consistently and significantly exceed the corresponding elements of P^2, P^3, $\cdots$. This indicates that (contrary to the assumption underlying a Markov process) a prolonged employment in a category decreases the probability of a move into another category. Accordingly the authors refine the model by assuming that the entire population consists of two strata, the "stayers" and the "movers." The stayers never move whereas the movers are subject to a random process of the type just described. If the relative sizes of the two strata are p and $q = 1 - p$, respectively, the present model predicts that $F_n \approx pI + qP^n$, where I is an identity matrix. However, this model actually overestimates the diagonal elements of F_n. The present model is a special case of an exceedingly flexible and useful model to which the authors call attention. Instead of assuming that the stayers never move, we may assume that each of the movements in each stratum are subject to a Markov process as described above with matrices P and Q, respectively. Now we should have $F_n = pP^n + qQ^n$. Instead of two strata one may consider a larger number of strata, thus attaining higher accuracies. (Further modifications are indicated in the last chapter.) Models of this type could be useful for learning theory even more than for labor movements (where probably the after-effect of the past history is so pronounced that higher-order Markov chains must be introduced).

Princeton University *William Feller*

*Cf. R. R. Bush and F. Mosteller, *Stochastic Models for Learning.* New York: Wiley, 1955.

PAUL E. MEEHL. *Clinical and Statistical Prediction.* Minneapolis, Minnesota: University of Minnesota Press, 1955, pp. x + 149.

In a little study I made recently of the interest and value structure of psychologists, the first principal component separates the laboratory from the clinic, the second separates the global-verbal from the analytic-psychometric approach to personality. It is hardly surprising, therefore, that the clinician and the statistician see things in different ways and have difficulty in communicating with one another. Meehl has attempted, in his present monograph, to achieve some reconciliation of these views and to build common ground between these groups.

Speaking as a psychometrician-statistician, I feel he has been quite successful. At least, there was little that I felt inclined to quarrel about in his presentation, and I feel I have a more sympathetic understanding of the clinician's activities as a result of my reading. I have not obtained the reactions of any thorough-going clinicians to find whether they felt equally satisfied.

There is not, as Meehl points out, one single clear issue as between the clinically oriented and the statistically oriented psychologist. Rather, there are a series of sub-problems. Thus, one issue concerns the value of psychometric as compared with non-psychometric *data*. A second concerns the efficiency of mechanical and non-mechanical ways of *combining* either type of data for purposes of prediction. One important distinction is between *structural statistics*, which aspire to provide a framework of constructs to describe the nature of the individual, and *validation statistics*, which undertake only to indicate the degree of association between prediction (whether stemming from scores or from clinical insights) and the course of subsequent events. A second, and perhaps the most central, contrast that Meehl makes is between the *context of discovery* and the *context of justification*.

A major part of the monograph is devoted to reviewing existing studies comparing the mechanical and non-mechanical modes of combining data for predictive purposes. Though the studies have many limitations, their general trend seems to definitely favor mechanical modes of combining, when the criterion to be predicted is some pre-established set of socially defined categories. These are such categories as level of academic grades, progress in recovery under therapy, or lapsing from grace during parole.

The primary predictive function for the clinician, as Meehl sees it, lies within the context of discovery in relation to the individual. That is, the distinctive contribution of the skilled clinician is that he can create the hypotheses that relate and apply general psychological principles to the uniqueness and complexity of the individual case. With what regularity these hypotheses are supported by later events remains a statistical problem. The need for tests of the accuracy of this hypothesizing is as great as the need to test the accuracy of predictions resulting from any psychometric device.

Teachers College, Columbia University *Robert L. Thorndike*